The Truth about Value and the Value of Truth

Johannes Petrus Smit

Dissertation presented for the Degree of Doctor of Philosophy (DPhil) at the University of Stellenbosch

Promoter: Professor AA van Niekerk

Co-promoter: Dr SA du Plessis

December 2003

Declaration

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work and has not previously in its entirety or in part been submitted at any other university for a degree.

26 November 2003

JP Smit Date

Abstract

In this thesis an attempt is made to sketch the outlines of a cognitive theory of ethics, i.e. a theory in which ethical statements are a subset of descriptive statements. It is argued that the good is a *quale*, and that this *quale* roughly corresponds to what is often referred to as "pleasure". If this conceptualisation of the good is correct, then the resulting ethical theory is a cognitive, egoist version of ethical hedonism. The thesis proceeds by relating this conceptualisation of the good to economic phenomena. An investigation is then made of the conditions under which the following of a boundedly rational rule is preferable to calculating which one of the possible options available to the actor to follow. It is argued that one such rule is that "truth" should serve as a norm of inquiry. Next the issue of "altruism" is considered. It is argued that our intuitions regarding what egoist action should be are radically untrustworthy. Considerations from evolutionary biology and game theory make it clear that an egoist actor might well be best advised to perform a number of actions that would normally be termed altruistic. The next topic concerns the relation between fact and value. Arguments that claim to undermine the distinction between fact and value are argued to be fallacious. It is also argued that the correct view of the relation between fact and value can help to clarify some of the problems surrounding the conceptualisation of "objectivity". The thesis ends by considering the gains that arise from adopting the position argued for.

Abstrak

In hierdie tesis word 'n poging aangewend om 'n kognitiwe teorie van etiek, m.a.w 'n teorie waarbinne etiese stellings 'n subspesie van deskriptiewe stellings is, daar te stel. Daar word geargumenteer dar "die goeie" 'n *quale* is, en dat hierdie *quale* rofweg dieselfde objek is as wat dikwels na verwys word as plesier. Indien hierdie siening van "die goeie" korrek is, dan impliseer dit die moontlikheid van 'n kognitiewe, egoïstiese weergawe van etiese hedonisme. Die tesis poog eerstens om die verhouding tussen hierdie siening van "die goeie" en ekonomise fenomene te verduidelik. Daarna word ondersoek ingestel na die kondisies waaronder die volg van 'n begrensde rasionele reël 'n beter opsie vir 'n akteur is as om an al die moontlike opsies te kyk en die beste te kies. Daar word geargumenteer dat die idee dat die "waarheid" die doel van ondersoek moet wees een so 'n reël is. Volgende word daar gekyk na die kwessie van altruïsme. Daar word geargumenteer dat ons intuïsies insake die aard van egoïstiese optrede radikaal onbetroubaar is. Sekere kwessies in evolusionêre biologie en spelteorie laat dit blyk dat 'n egoïstiese akteur waarskynlik verskeie oënskynlik "altruïstiese" aksies behoort uit te voer. Die volgende kwessie wat bespreek word is die verhouding tussen feite en waardes. Daar word geargumenteer dat pogings om hierdie onderskeid te ondermyn nie suksesvol is nie. Daar word verder geargumenteer dat die korrekte siening insake hierdie verhouding sekere probleme insake die verstaan van "obkjektiwiteit" kan ophelder. Die tesis eindig deur die voordele wat uit spruit uit die aanvaarding van die posisie wat hier voor geargumenteer word.

The financial assistance of the National research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the National Research Foundation.

Acknowledgements

I want to thank my supervisor, Prof. AA van Niekerk, for his guidance, and tolerance of ideas very different from his own. I wish to thank my co-supervisor, Stan du Plessis, for keeping some patent absurdities out of this thesis. I wish to than Prof WL van der Merwe for his encouragement and close reading of this thesis. In addition I would like to thank Prof. Don Ross for his detailed and constructive criticism, and tolerance of ideas *profoundly* different from his own. I also wish to thank friends and family who tolerated various rantings on the topics contained within. I would also like to thank the NRF and the Harry Crossley Foundation for their financial assistance.

Opgedra aan my ouers

Table of Contents

Introduction

1. What is Ethics?

There is a sense in which words matter, and a sense in which they do not. Which word is chosen to refer to a concept is unimportant. It is a convention that can be changed without affecting anything, cognitively speaking. What is important is that its usage be intelligible and that it be consistently applied.

Consistently with the above, Karl Popper often warned against the essentialist urge to ask "what is ..?" questions[1]. Such questions serve no purpose. If two people use the same term to refer to different things there exists no substantial disagreement between them. Both parties just need to realise that they are simply talking about different things, and any possible reasonable dispute should be resolved.

This thesis concerns, in part, "ethics", here defined as "the attempt to answer the question 'how should one act?'. The answer that will be defended can be described as hedonist egoism. In other words, it will be claimed that one should seek to maximise one's own pleasure, and that this answers the fundamental question of ethics.

This answer might well appear strange or counter-intuitive. This, admittedly, is the case. This thesis will try to make it sound a lot less strange and counter-intuitive. For now I wish to address the single objection that I have, both within formal contexts and informally, most encountered with reference to the idea of egoist hedonism. This is that it "is not ethics".

The best reply to such an objection is to state that "one should maximise one's pleasure" is definitely an answer to the question "how should one act?". This is the sense of the term "ethics" used in this thesis, and it is a sense that is as well established historically as any other philosophical term. If the questioner has any other sense of the word "ethics" in mind it might well be that this thesis does not concern "ethics", as the questioner uses the term. There are lots of things that this

[1] See Popper (1976: 18-31) for a defence of nominalism against the kind of essentialism that leads to "what is..?" questions.

thesis does not concern. The fact that the questioner uses the letters "e-t-h-i-c-s" to refer to one of them does not make it any more relevant than any of the others.

The word "ethics" is sometimes used in ways other than to refer to "how should one act?". Provided that these uses are kept separate, this would not cause any trouble. But, when it comes to a matter as contentious as answering this question it would seem that nothing is straightforward. One alternative definition of ethics is not to define it as "how should one act?", but in terms of one particular answer to this question, i.e. in terms of a value-judgement. In this manner some would define "ethical action" in terms of its supposed opposite, i.e. self-interested action. Here a specific *answer* to the question "how should one act?", namely "one should take others in to consideration", is used to define the idea of ethics. In other words the definition of ethics, in this case, already includes a value-judgement. There is nothing necessarily wrong with such a definition. But it needs to be remembered that, on such a definition, pronouncing egoism to be unethical does not tell us anything. Or, rather it does not convey any insight regarding egoism or ethics, rather it just tells us how someone has decided to use a word[2]. If the person intends the statement to serve as a condemnation of egoism then he is engaged in sophistry, and trying to do by definition what needs to be done by logic.

There are many possible different definitions of ethics that define ethics in terms of a specific value-judgement regarding "how should one act?", as opposed to simply defining it in terms of "how should one act?" Peter Singer defines "ethics" in terms of "justifiability" (1993: 10-12). This seems to reflect his value-judgement that people should act in such a way that their actions can be "justified", in his sense. He realises however, that his *definition* does not allow him to condemn those who do not care

[2] This type of argument seems to be found in Cilliers (1998: 139). He states that to behave ethically means not to follow rules blindly – to merely calculate – but to follow them "responsibly". If this is taken as a value-judgement against blindly following rules, it would make sense. But the author writes that this is what it *means* to behave ethically. This is a strange way to state an argument. For, whether this is really what "ethical" means is a matter of arbitrary definition. Even if it is true, it does not give *any* rational support for the idea that one should not follow rules blindly. The author also discusses the possibility of following a universal set of rules (137), and wonders whether such action can be called "ethical". He proceeds to state that "what is at stake here is the very meaning of the word ethics" (137). The argument seems to concern whether one should try to follow (possibly illusory) rules or not. But it is not stated this way, rather it is written as if it is about the meaning of the word. Writing the argument in this form, I would contend, makes it easier to fool the unwary into thinking that a factual, rather than

about whether their actions are justifiable. Rather, if one defines "ethics" in terms of a given value-judgement, this just raises the question as to the judgement's validity. He acknowledges this by asking the question "Why be ethical?" (1993: 314-335). This is not a question that would be *possible* if one defined ethics as "how should one act?". Singer is conscientious enough to acknowledge that, given his definition of ethics, this becomes an intelligible question.

Ethics can, and has, been defined in terms of "justifiability", "universalisability", "one's relation to the other", etc. These definitions all seem to reflect the judgement that the specific element in terms of which ethics is being defined is ethically desirable. These definitions, however, all make possible the question as to whether these things really are desirable. As long as the author is aware of this, like Singer, this should not lead to unnecessary disputes. But I would contend that definitions like the above can be harmful to the understanding. If things are pronounced "unethical", but this is just the result of a definition, it is easy for the unwary mind to be taken in and to think that these judgements have some rational support behind them. Meanwhile it is possible that all that really happened is that the author declared actions she doesn't like to be unethical by definition. In other words, it is unclear what uses such definitions can have apart from the rhetorical.

There is also something strange about using the same term to cover a field of study and a substantive answer within this field of study. This is what is done when "ethics" can mean both "how should one act?" and "[the study of] the kind of person one must become in order to develop a non-violative relation to the Other" (Cornell, 1992: 13). Here ethics refers both to the question "how should one act?" and the answer "one should be in a non-violative relation to the other". The above is a bit like defining physics as "the search for regularities in nature" and as "the investigation of the implications of Newton's work". In other words, in both cases a field of study is conflated with an influential view in this field of study.

I conclude that, while definitions of ethics that explicitly contain a value-judgement are not necessarily bad, it is far from clear what useful purpose they have. It was

---

value-issue is at stake. In accordance with Popper above it must be stated that, if "the very meaning of the word ethics is at stake", then, nothing at all is at stake.

explained earlier what *unreasonable* purpose they can have, namely giving the illusion of reason where none exists.

Above it was firstly argued that the existence of definitions of ethics other than "how should one act?" are irrelevant to this thesis. This is because, as with all "what is…? - questions" there is no such thing as the "correct" answer to the question "what is ethics?". There are simply a number of definitions of ethics, and there need be no discord between them. It was also suggested that existence of definitions other than the standard "how should one act?" serve no useful purpose, but can only be misleading tools of rhetoric.

2. Plan of the Work

This thesis wishes to defend a position that can be described as cognitive, egoist, ethical hedonism. It is cognitive since it states that values are a subset of facts, not something radically discontinuous with facts. It is egoist and hedonist, for reasons mentioned already. The position is called *ethical* hedonism to distinguish it from *psychological* hedonism. Ethical hedonism is the view that one *should* seek pleasure, psychological hedonism the view that people *do* seek pleasure.

The above view might appear extremely peculiar. But in truth it rests on mixing together three basic ideas, all of which have, historically, a degree of philosophical respectability.

The first is hedonism. Its respectability might not be what it once was, but it still has its defenders. I would also argue that the version of hedonism developed here allows for certain traditional objections to hedonism to be overcome.

The second concerns the question "Do we call it good because we like it, or do we like it because it is good?". If our pursuit of pleasure is understood to be the result of pleasure being "good", rather than because we simply want pleasure, then this allows for cognitive statements regarding value.

The third concerns *qualia*. If pleasure is accepted as standard, then it becomes clear that what makes pleasure capable of being a standard of action, i.e. what makes it "good", is irrevocably tied to the individual experiencing it. Simply put, no-one's pleasure can matter to me in the same immediate way that my own does. This observation is used to argue that pleasure (or "good) is a *quale*, and that hedonism inevitably leads to egoism.

Chapter 1 will concern the main theoretical basis of this thesis. It will be argued that one thing can be found which is "self-justifying" or good "in-itself" – pleasure. It will be argued that accepting this inevitably leads to cognitive, egoist, ethical hedonism.

Chapter 2 will look at the various concepts involved in this position and try to clarify any difficulties that their usage might raise. It will also look at some historically influential objections to the ideas being defended.

Chapter 3 will take this account of human action and relate it to economics. Economics is a discipline that has a rich tradition of theorising regarding value. Therefore it is interesting to relate this view of value to what has happened, over the last few hundred years, in economics. It also briefly relates the theory of value to economic phenomena proper. This is done in order to show that it can make sense of these phenomena.

Chapter 4 is the first of the chapters that try to give content to the injunction "act so as to maximise pleasure". It will be argued that an actor with limited information and constrained computational capacities will best maximise pleasure by being a rule-following, rather than a calculating creature. It will also be argued that the design history of the mind makes it almost impossible for human beings to be calculating creatures. In other words, "satisficing" trumps "rational choice" where human beings are concerned.

Chapter 5 will attempt to determine the content of one such rule that a pleasure-maximiser should follow. A qualified version of the argument that "truth" should be a norm of inquiry will be argued for.

Chapter 6 will look at the objection that is most often levelled at egoist theories of ethics. This objection is that the existence and desirability of altruism shows egoism to be false, both as a descriptive and normative theory. The chapter will argue that our intuitions as to what egoist action would look like are radically untrustworthy. It will be argued that we have good reason to suppose that a succesful egoist would commit any number of actions that would not normally be considered to be nice. The possibly counter-intuitive nature as to what would constitute egoist action can be shown by considering the evolution of preferences, the structure of social interaction, the bounded rationality of the actor and the possibility of cognitive mistakes.

Chapter 7 concerns more abstract matters. It looks at some further epistemological implications of the view of value that is defended in this thesis. It will be argued that certain confusions and problems regarding "objectivity" and the "fact/value" dichotomy can be avoided by adopting the view defended here.

Chapter 8 is the final chapter. It will clarify some issues that arose in the thesis, without explicitly being addressed.

It remains to clarify an important presupposition of this thesis, and to give some background to the problem that will be discussed.

3. A Note on Folk Psychology as Presupposition of this Thesis

Neurath's dictum to the effect that, if one wishes to rebuild a ship at sea one needs some place to stand, while not universally acknowledged as a correct characterisation of philosophy as such, definitely does apply to the writing of most particular works of philosophy. Issues in philosophy tend to be related to one another in such a way that it is hard to tackle a specific topic without making some assumptions concerning related matters. This is also true of this thesis. The most important commitment of this thesis that will not be extensively argued for concerns folk psychology.

Folk psychology is generally taken to be a conceptual framework in terms of which actions can be explained or predicted. This conceptual framework includes terms like "belief, desire, pain, pleasure, love, hate, joy, fear", etc. (Churchland, 1994: 308). In

other words, folk psychology concerns a certain view of action that characterises the actor as a being with propositional attitudes like beliefs and desires, and who acts upon these beliefs, often to satisfy these desires.

The most controversial issue surrounding folk psychology concerns the status of the objects of folk psychology, i.e. the "beliefs", "desires", "intentions", etc. All would acknowledge that we cannot really get by without folk psychology, but opinions differ as to why this is the case. Some believe that folk psychology is a useful way of coping with the world, but do not take the apparent ontological commitments of folk psychology seriously. Others argue for the approximate ontological adequacy of folk psychology[3].

This thesis will use the categories of folk psychology, and does partly depend on the categories of folk psychology being ontologically respectable. While some of the claims made might well have a truth-preserving interpretation in terms of a view that does not view folk psychology as ontologically adequate, a lot of the central claims of this thesis become unintelligible unless the categories of folk psychology, and specifically *qualia,* are assumed.

It is often objected that the folk psychological categories should be examined in the same manner as those of any other scientific theory. It will then be found that beliefs, intentions, desires, etc. are like other theoretical entities, for instance phlogiston, that have no explanatory value regarding the nature of reality. By an application of Occam's razor it is then best to get rid of them *qua* understanding the nature of the world. Folk psychology would still be an indispensable tool for coping with the world, but would not be understood as having ontological import.

Something like the above position has been argued for by a number of authors[4], and it is not hard to guess what the objection would be. It is objected that beliefs and desires are not like phlogiston in that their existence is not inferred in order for them to have

[3] A good introduction to the issues can be found in Guttenplan's *Companion to the Philosophy of Mind* (1994). It is indicative of the degree of controversy surrounding folk psychology that two rather divergent entries are listed under the heading "folk psychology".

some explanatory function[5]. We experience ourselves as having intentions, we do not observe phenomena and then come up with the theory of having intentions. In other words, intentions belong to the primary data that need to be explained, not to the explanatory structure we erect in order to explain data. To this it will then be objected that we only ever observe behaviour, which is then interpreted in terms of intentional concepts.

The above problem is closely related to the difference between "objectivity", understood as the so-called third person view of the world, and "subjectivity", understood as the so-called first person view. How we think about *qualia* and folk psychology seems to largely be a function of whether we start with the third person view or the first person view. A large amount has been written about the above topic, especially after Nagel's famous "What is it like to be a bat?" (1974). It is not a topic to which I intend to contribute here.

The problem with addressing the topic of folk psychology is that any number of controversial issues are directly related to it. For instance, consider the issue concerning which is primary: the first-person view or third-person view. There is a long tradition in philosophy from idealism to phenomenalism to positivist instrumentalism that seems to accept some version of the claim that sense-data constitute "data", which then must be accounted for by theoretical entities like physical objects. This type of view is less common in contemporary philosophy, no doubt because of our current understanding of what the prestige of science consists in, and also some specific work in philosophy, like Quine's attack against what he termed the second dogma of empiricism (Quine, 1965). Launching into such a complex topic, one that has been discussed and resurrected any number of times throughout the history of philosophy, is not really feasible in terms of the main goals of this thesis. And the problems would not end there, for the question as to which is "primary" is not a question that could be usefully discussed without also bringing in any number of other issues in the philosophy of language and philosophy of mind. And treating some

---

[4] See, for instance, Dennett's *The Intentional Stance* (1987) for a version of the above position. It includes an enlightening comparison between his view and the view of other writers who have addressed the topic (1987: 339-350).

[5] See, for example, Searle (1992: 59): "We do not p*ostulate* beliefs and desires to account for anything. We simply experience conscious beliefs and desires.".

other philosophical assumptions as primary in order to give a defence of folk psychology would not, in the context of this thesis, be that great an advance over simply declaring that this thesis presupposes the ontological adequacy of folk psychology[6].

If the above view is mistaken, then a lot of what is interesting in this thesis would disappear. It would not, however, affect everything that will be said. Some of the topics in this thesis, especially as related to economics, decision theory and epistemology can survive the death of folk psychology unscathed. But the central claim of this thesis, i.e. that the good is a *quale* we should pursue, would only make sense if the folk psychological categories are accepted[7].

4. A Note on the Motivation for this Work and History of the Problem

There is one final point I wish to add. The research that led to the writing of this thesis centred on the question of whether a cognitive ethics is possible. The egoist or hedonist claims made within might well be more surprising, but I view the merit of the present position in that it attempts to find a cognitive theory of ethics.

Ever since Hume formulated the "is"-"ought" problem, philosophers have had difficulty defending the rationality of value-judgements. If one accepts a basic empiricist commitment, then Hume's criticism is devastating. Hume's problem can only be overcome by changing what counts as rationality, i.e. deviating from the basic framework of empiricism. Another option would be to accept two different, and incommensurate, standards of rationality for factual knowledge and ethics. I am extremely sceptical of both these options. Yet it seems that, if one wishes to have one standard of rationality, and one wishes this to conform to basic empiricist standards, there can be no matter of fact with reference to value-judgements.

---

[6] Simply nailing one's colour to the mast is not an uncommon strategy to take. For instance Dennett, one of the great opponents of allowing the objects of folk psychology to be part of reality, occasionally asserts that he will assume the third-person point of view as it matches his "intuitions", or that this choice is based on a "tactical hunch" (1987: 7), as opposed to a conclusive argument.

[7] In accordance with this adoption of the first-person perspective, an internalist account of mental content will also be treated as a presupposition of this thesis. In other words any reference to mental

The early part of this century, mostly due to the early Wittgenstein and the positivists, underlined this claim. Sticking to empiricism, and not wishing to have another, incommensurate set of standards for rational belief, they could not find a place for value in the world. Of course, this was in no way particular to the positivists. For example, the existentialists could also find no factual difference between right and wrong. But the position was most closely associated with the positivists. They were also the group most maligned for holding these positions.

I would contend that, with regard to cognitivity, the position hasn't changed much. After the positivists, and under the influence of the later Wittgenstein, Austin and others, philosophers started inquiring after the function of the language that an earlier generation had called nonsense. Here, instead of calling ethical language nonsense, they started viewing ethics as non-cognitive, yet meaningful. Ayer had, at this point, already put forth the view that ethics was a matter of expressing wishes. These "attitudinist" type of positions were a way of giving ethical language meaning, without letting it refer to reality. In due course philosophers were arguing about whether ethical language is the giving of commands, the expression of wishes, prescriptions, etc. The most influential theory of this time was probably HM Hare's "universal prescriptivism", whereby to state an ethical rule is to prescribe a course of action for all.

All the above theories, however, agree that ethical language has no cognitive meaning. The prevalence of this idea has persisted till today. A survey of different ethical positions on the philosophical landscape would probably turn up a bewildering array of possibilities, but very few would view ethics as cognitive.

I think the history of 20th century ethics in the analytic tradition is basically the history of trying to cling to empiricism, while trying not to totally condemn ethics to the wastebasket. This entailed giving up the idea of cognitivity, but keeping the idea of meaningfulness.

---

content should be read as referring to the narrow conception of mental content, and it is being assumed that "narrow content" is instantiated in our world.

It is within this context that the present thesis should be seen as an attempt to solve the problem in a slightly different manner. It is contended here that empiricism can be kept and that the idea of only one standard of knowledge can be kept. Value-judgements should be seen as being true or false in virtue of extra-linguistic entities, the nature of which will be discussed in chapter 1. In other words value-judgements can be squeezed to fit into even the very narrow standards of acceptability that the positivists had. One need not fundamentally change one's epistemological views if one wishes to keep cognitivity.

This thesis is an attempt to present one possibility regarding what one does need to believe in order to achieve cognitivity, and an argument for the truth of these beliefs. The basic claim of which I wish to convince the reader is that, what has always been called "pleasure", is actually "the good". Chapter 1 will make this argument.

Chapter 1: The Object of Ethics

1. Introduction

This thesis will argue that value-judgements are a delineable *subspecies* of factual judgements, rather than something radically discontinuous with it. The continuity between fact and value vests in the claim that both factual and normative discourse can be true or false in virtue of a representation-independent matter of fact, i.e. that truth-conditions can be assigned for both. The investigation of such truth-conditions for normative discourse forms the main topic of this thesis.

This chapter is concerned with the nature of the extra-linguistic entity that serves as justification for the claim that value-judgements are true or false. My argument will firstly establish the conditions required for an entity to affect the truth-value of a value-judgement. It will be argued that there is an entity that meets these conditions, and a brief characterisation of this entity will be given. The second chapter will deal with some difficulties regarding the conceptualisation of this entity, as well as drawing some important implications from this determination.

2. Definitions of Fact, Value and Ethics

A "fact" will be defined as an extra-linguistic entity, in virtue of which the truth-value of a proposition is fixed. This definition is deliberately vacuous, since the claims made in this thesis can be consistent with more than one conception of "fact" or "truth". The only constraint that it is necessary to specify at this point is that the conception of fact must be conceptually distinct from the conception of "value". It would be inconsistent with a crude pragmatism that takes "usefulness of belief" to be the only *possible* constraint on belief-formation[8]. Any constraint, or set of constraints,

[8] It seems unlikely that anyone has ever consistently held this position. Certainly James would reject it as "the usual slander, repeated to satiety by my critics" (1978: 147). Rather he would say that correspondence, coherence and value are the necessary components of "pragmatic truth". Thayer's introduction to the 1975 edition of *The Meaning of Truth* interprets James in this way. Roughly this interpretation is followed by Hallberg (1997: 205-223). Ayer, in the introduction to a joint edition of *Pragmatism* and *The Meaning of Truth* (1978), interprets James as dismissing the notion of fact, and hence as propounding something similar to the type of crude pragmatism defined above. It seems equally possible that James vacillated between these two positions. See chapter 7 for a discussion of James' position.

on belief-formation over and above "usefulness" can serve as an adequate conception of "truth", and consequently "facthood".

"Value" will be defined as the object of value-judgements. In other words "value" is being defined as what is referred to by the "value-term" in a value-judgement. A "value-judgement" is any judgement regarding the irreducible[9] goodness, badness, rightness, etc. of something. Note that nothing is being pre-judged by this definition of value as "object". Non-cognitivists might allow this definition, but merely maintain that value, as defined, does not exist.

A large part of this thesis will concern "ethics". "Ethics" is here defined as the attempt to answer the question "How should I/we act?". This attempt must ultimately contain value-judgements, as defined above, that seek to guide action. Hence any statement that includes an irreducible value-term and seeks to guide action qualifies as ethical.

The first issue to be considered concerns the conditions for the cognitivity of value-judgements. This issue is nicely brought into focus by the early Wittgenstein's famous denial of the cognitivity of value-judgements. This will be examined in order to see what exactly he denied when pronouncing ethics, etc. "unsayable".

3. The Conditions for Cognitivity

3.1. The *Tractatus*' epistemology

In the *Tractatus* Wittgenstein is concerned with trying to explicate the limits of what can meaningfully be said[10]. Propositions are meaningful if they refer to facts; this implies that the limits of meaning are fixed by the limits of what can possibly be referred to. Ultimately these referents are the so-called "atomic facts"[11], these are constituted by "objects"[12].

---

[9] "Irreducible" is meant to exclude any uses of value-terms (good, bad, right, etc.) that can be fully analysed into non-value-terms, as well as the equivalent "conditional should".

[10] "The book will, therefore, draw a limit to… the expression of thoughts…". (1960: 27). (Note: All references are to the 1960 edition of the Ogden translation).

[11] "2. What is the case, the fact, is the existence of atomic facts."

The essence of Wittgenstein's conception of knowledge is that a proposition is only meaningful if there exists objects that can ultimately ground its meaning. If this is not the case, then a speaker is literally "talking about nothing". The idea of judging such statements about nothing as either true or false is preposterous; they are simply meaningless.

In this manner the requirement of cognitivity becomes the requirement for the existence of a Wittgensteinian "object"[13]. Wittgenstein never specifies a *numerus clausus* of objects, neither does he commit himself to any traditional ontological doctrine regarding "what is really out there". He does assert that "natural science" contains real propositions (6.53). This amounts to saying that there are some representation-independent entities that make the propositions of natural science true or false.

2.2. The Object of Ethics

The converse is then true when interpreting his remarks on ethics. His statement that there can be no "ethical propositions" (6.42) is equivalent to the statement that no object of ethics can be identified. This is a (substantive) *ontological* claim, a claim that is, in principle, independent of his (formal) referential theory of meaning. If Wittgenstein had supposed that such an object of ethics could be identified, then a description of this object would have fitted seamlessly into the *Tractatus*. Instead of the famous "Whereof one cannot speak, thereof one must be silent", statement seven could have contained an identification of the object of ethics and, presumably, an explanation of the consequences of this identification[14]. In this manner ethics would have become another "natural science" on a par with other natural sciences.

---

[12] "2.01. An atomic fact is a combination of objects (entities, things)."

[13] No two *Tractatus* commentators seem to agree on exactly what would constitute a Wittgensteinian "object". (See, for a discussion of various positions and another suggested interpretation, Goddard and Judge (1982)). The situation is exacerbated by the fact that Wittgenstein explicitly repudiated (1994: 21e) his characterisation of objects as "simple" (2.01). This difficulty is glossed over in this thesis on the assumption that thinking in terms of "propositions" and "what they refer to" is not fundamentally misguided, i.e. that the difficulty lies with the characterisation of "objects referred to", and not the very idea of "objects referred to".

[14] After implicitly denying the existence of an "object" of ethics, Wittgenstein does add some cryptic remarks, the most important being that ethics is "transcendental" (6.421), and that "[t]here is, indeed, the inexpressible" (6.522). These remarks will not be considered, rather his argument will be challenged at a point before the need for such remarks arise. Such treatment of the *Tractatus* is closely

Wittgenstein denied that such an object of ethics exists, but did give a brief explanation of what would be required of such an object. He writes that a sufficient answer to an ethical question cannot merely refer to the consequences of a given action, but must be capable of stating its case only with reference to the action itself (6.422)[15]. This appears to be a restatement of the old point that justifications in terms of instrumental value lead to an infinite regress if some object of inherent value is not identified. Any candidate for the role of object of ethics needs to have, in some sense, inherent value, or, since it amounts to the same thing, be *self-justifying*. This self-justifying element would have to be of such a nature that the idea of further justification of an action does not arise and the infinite regress ends.

It should by now be clear that Wittgenstein's denial of the cognitivity of ethics amounts to the claim that no object can be identified that has inherent value. Or, alternatively phrased, that no object can be found which is self-justifying in a sense which allows the regress of justification to end.

To deny that something can *literally* have "inherent value" hardly amounts to a controversial philosophical position. It might well seem to be no more than a denial of the most essentialist, absolutist and metaphysically dubious position imaginable. Indeed, one need not have any particular regard for the *Tractatus* to take this position. Mackie, who has no particular positivist sympathies, dismisses the idea of a cognitive ethics. His reasoning for this position is very similar to that of Wittgenstein, as will be shown below.

3.3. Mackie's "Argument from Queerness"

Mackie explicitly disavows any belief in the existence of a self-justifying element, yet he states that belief in the existence of such an element is a constitutive part of moral discourse (1977: 48). Since such an element does not exist, moral discourse must take

---

aligned to the spirit of Ramsey's famous "[b]ut what we can't say we can't say, and we can't whistle it either", or Neurath's "[o]ne must indeed be silent, but not *about* anything" (Ayer, 1985: 32).

the form of an "error theory" (1977: 48-49). His "argument from queerness" gives a good indication of what would be needed for ethics to be cognitive:

> An objective good would be sought by anyone who was acquainted with it, not because of any contingent fact that this person, or every person, is so constituted that he desires this end, but because the end has to-be-pursuedness somehow built into it. Similarly, if there were objective principles of right and wrong, any wrong (possible) course of action would have not-to-be-doneness somehow built into it (1977: 40).

Mackie argues that the idea of "to-be-doneness" being built into an element of experience[16] is exceedingly peculiar, so peculiar that it is scarcely imaginable what element could possibly satisfy such a requirement. This is then used to dismiss the idea of the existence of to-be-doneness, and hence the possibility of a cognitive ethics (1977: 49).

While both Wittgenstein and Mackie agree that ethics cannot be cognitive, they do seem to share roughly the same idea about what would be required for a cognitive ethics. For an entity to have Mackie's to-be-doneness would be for it to necessarily require action without reference to some goal outside itself, i.e. it would need to be self-justifying, or have inherent value, etc. These phrases amount to different ways of affirming the same thing, namely that it can end the regress of justification in a non-arbitrary manner.

Another way to say this would be to state that the entity in question is good "in-itself", or simply "good". Note that this entity would not be *evaluated* as good, but, in some sense, simply *be* good. For if it is a matter of being evaluated as good, then it might be asked whether the standard of judgement used in this valuation is good, in which case the same regress occurs as was referred to above. This inquiry can only

---

[15] "6.422....There must be some sort of ethical reward and ethical punishment, but this must lie in the action itself".

[16] Mackie and Wittgenstein's dismissal of an object of ethics is in agreement with Hume's famous denial of the idea that such a peculiar quality ("vice") is an element of experience: "Take any action

end if the standard of judgement is "goodness itself", and hence "good" somehow functioned referentially.

Below it will be argued that Mackie's "metaphysically peculiar" (1977: 49) element does exist, and that it can serve to make ethics cognitive in a way that satisfies even the strict conditions of meaning in Wittgenstein's *Tractatus*. This will be done by looking at the writings of Schlick, one of the few, along with Kurt Baier, of the positivist admirers of the *Tractatus* who wrote substantially about ethics.

4. Cognitive Ethics and Schlick's "Tone of Experience"

3.1. The "Tone of Experience"

Is there an "object" of the good, i.e. does there exist a legitimate referential use of the term "good"? Schlick's characterisation of hedonism in *Problems of Ethics* seems to suggest this possibility:

> Every idea, every content of our consciousness, as we learn from experience, possesses a certain *tone*. And this has the consequence that the content in question is not something completely neutral, or indifferent, but is somehow characterised as agreeable or disagreeable, attractive or repellent, joyful or painful, pleasant or unpleasant.... The essence of these feelings is of course indescribable – every simple experience is of course beyond description – and one can only make clear what is meant by appropriate indications (1962: 37).

Schlick maintains that the *content* of part of our experience "is not completely neutral", and this quality of "non-neutralness" is "indescribable". He further states that the pursuit of this quality constitutes the "law of human motivation" (1962: 36-40). In doing so he joins a long line of philosophers who have advocated some form of psychological hedonism.

---

allowed to be vicious: willful 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*..."(Hume, 1969: 520).

The phrase "psychological hedonism" is used to refer to thinkers who think that man ultimately *is* controlled by pleasure or pain. The phrase "ethical hedonism" is usually used to denote those who think that man *should* strive after pleasure or pain. Ethical hedonism is usually advocated by those who also believe in psychological hedonism, which gives rise to the question as to whether these two positions can be made consistent and what the ethical *should* could possibly mean in this context[17].

In this thesis the relation between psychological hedonism and ethical hedonism will be, in an important sense, reversed. This chapter will provide an argument for ethical hedonism. In chapter 2 it will be shown why a version of psychological hedonism follows from an acceptance of ethical hedonism.

This is diametrically opposed to the position that Schlick ended up taking. Ultimately he states that "ethics" can only ever be a branch of psychology (1962: 29), since it can only be concerned with what people do and not what they should do[18]. The "should" can be used hypothetically, i.e. with reference to an already assumed goal, but the idea of the justification of a final ground is "senseless" (1962: 18).

But, if ethics is considered to be concerned with the question "how should one act?", and this "should" is not the unproblematic "hypothetical (conditional) should", then it is precisely this justification of a final ground that is sought. What is sought is an element that is self-justifying, that simply is to-be-done and that ends the infinite regress of justification in a non-arbitrary manner. Above it was shown that Schlick explicitly disavows the possibility of the justification of a final ground. In similar vein Wittgenstein said that "[o]ught in itself is nonsensical" (Waismann, 1979: 118). Hence Schlick's *Problems of Ethics* does not, fundamentally, constitute an exception to the association between positivism and non-cognitive theories of ethical language.

---

[17] Bentham's doctrine has had to face this problem. For a discussion and one proposed resolution of the problem, see Hart and Burns' Introduction to the 1982 edition of *The Principles of Morals and Legislation*.

[18] This is consistent with Wittgenstein's assertion that "...the will as a phenomenon is of interest only to psychology"(6.432).

Psychological hedonism is perfectly compatible with believing ethics to be non-cognitive or expressive[19].

Schlick's denial of the possibility of the justification of a final ground of ethics amounts to the denial of a legitimate referential use of "good". But his formulation of psychological hedonism does point the way to a formulation of hedonism that includes a legitimate referential use of "good" – and an object of ethics.

3.2. The Referential Use of "Good"

Consider the synonyms that Schlick provides for his indescribable notion of the "tone" of experience. He mentions, among others, "agreeable", "joyful", and "pleasant". It will surely be uncontroversial to assert that there are certain instances where we use a value-term to refer to these experiences; this is the use of "good" as included in the phrase "feeling good". All manner of "pleasurable" experiences are commonly referred to in this manner. We will say that something "tastes good", "sounds good", "smells good", etc. No doubt there are some instances where we use these phrases to refer to something other than the pleasure evoked by these experiences. But, it does seem reasonably clear that there are instances where "it feels good" and "it feels pleasant" are synonymous. Hence there is an established usage of the term "good" that refers to the same element of our experience that Schlick referred to as the "tone" of our experience. In other words, there is an established usage of the term "good" where it functions referentially.

This chapter will defend the idea that this usage of "feeling good" constitutes a legitimate referential use of the term "good", and that this is the fundamental clue towards developing a satisfactory theory of cognitive ethics. It will be argued that the term "good", as it is used in "feeling good", provides the fundamental atoms or building blocks for the understanding of axiological phenomena. In other words the basic hedonist claim will be made that all axiological phenomena (ethics, aesthetics, economics) can be explained as deriving from the basic idea of "feeling good". It will

---

[19] Schlick seems to end up with an "attitudinist/expressive" theory of ethical language, interpreting ethical language as the expression of desires: "Everyone knows [that] 'I ought to do something' never means anything but 'Someone wants me to do it.'"(1962: 110).

be argued that this usage allows Wittgenstein's claim that there is no value *in* the world (6.41) to be challenged.

The immediate objection to such a position based on the common usage of "feels good" as synonymous to "feels pleasant" is that it overestimates the importance of this usage. It could perhaps just be a linguistic curiosity that is unrelated to the idea of "the Good". If there existed an established usage of the term "seeing a car" that was commonly taken as synonymous to "seeing a good", and nothing definitively stops us from establishing such a convention, nobody would be much impressed by an argument that claimed cars to be the object of ethics. Here the term "good" as a synonym for "car" has nothing to do with "the Good", the term synonymous to "car" just happens to be a homonym for "good".

Is "feeling good" unrelated to "the Good" in the same manner as the example constructed above? The answer is no. The argument for relating "feeling good" to "the Good" is important in that it constitutes an argument for ethical hedonism, and clarifies the relation between ethical hedonism and psychological hedonism.

3.3. "Feeling Good" and "The Good"

How can "feeling good" be related to "the Good"? It will surely be conceded that, however undeveloped or even nonsensical our conception of "the Good" is, it is inextricably linked to the idea of something being self-justifying. Ethics has commonly been thought to concern the very idea of "justification"[20], with the qualification that such justification is of a specifically "ethical" type. When an action is thought to be justified in the specifically ethical sense, then it is not merely supposed that the action accomplishes a certain goal. It is also being implied that this was a "worthy" goal of action, i.e. a "good" goal. Or, in other words, to call something "good" in the "specifically ethical" sense is another way of saying that it is self-justifying.

---

[20] Some philosophers have claimed that the idea of "justification" is *the* element of ethics most characteristic of it. For instance, Singer: "The notion of living according to ethical standards is tied up with the notion of defending the way one is living, of giving a reason for it, of justifying it." (1993: 10). Also compare Rescher's provisional definition of "a value": "A value represents a slogan capable of providing for the rationalisation of action..." (1969: 9).

If it is agreed that our conception of something being "ethically good" is inseparable from it being self-justifying, then this makes it possible to relate "the Good" to "feeling good". For, if it can be shown that "feeling good" is, under certain circumstances, self-justifying, then this shows that "feeling good" and "the Good" are related by more than linguistic accident. It will firstly be argued that we intuitively recognise "feeling good" to be self-justifying. Then it will be shown why this implies ethical hedonism and not psychological hedonism.

3.4. Argument 1: "Feeling Good" is Self-Justifying

All of us have friends who have habits or commit actions that strike us as strange, even bizarre (and, probably, *vice versa*). Most of us have accused others – or been accused – of enjoying awful music, eating sickening food or wasting time in pointless pursuits. The problem with these situations is that, to a bystander, these actions might simply make no sense. Whether it is listening to industrial music or bubblegum-pop, playing golf or bungee jumping, there will always be people willing to swear that these activities are wonderful and others who are extremely puzzled by these commendations.

Consider two people asked to justify such an action. Suppose that two people are asked why they play golf. In such a case the one person might well reply that playing golf gives her great pleasure, that nothing feels quite as "good" as the feeling of a well-struck shot.

Imagine, however, that the second person states that she plays golf because a friend also plays. In such a case we might well want to ask why the second person lets her choice be controlled by the opinion of a friend. If the person refuses to answer, and simply states that the goal of her actions is to imitate her friend, we are unlikely to accept this as a sufficient answer. If the person keeps insisting that she simply does what her friend does, and that there is no reason[21] beyond that, it might be suspected that she is not being *bona fide* in answering the question.

---

[21] "Reason" is here meant to indicate "goal". There might well be a "reason", in the sense of a prior event that caused the person to imitate her friend, but that is a different matter. Here the question of "reason" is the question "what goal are you trying to achieve?".

But if the first person is asked *why* pleasure/"feels good" is the goal of action, she might well consider this a puzzling question. She might answer that she plays golf because it gives her pleasure, and that there simply is no further motive. If the question is repeated, and it is asked why "pleasure" was a goal of action, she will think the questioner has misunderstood her answer. Here it will be the questioner who will be suspected of being *mala fide*, for how can anyone who knows what pleasure feels like not understand why it can be a sufficient goal of action?

It might be suspected that the person is not being quite truthful when she states that she derives pleasure from playing golf, it might also be difficult to imagine that the activity can give rise to pleasure. But, if it is accepted that the person does derive some enjoyment from the activity, their action gains a certain *intelligibility*. Simply put, the action now makes sense in a way that it didn't when the goal of it was unstated. This intelligibility is not only the result of a goal of action being specified, for then the answer "because my friend plays golf" would have given the action the same intelligibility. Rather the goal needs to be a *sufficient* or self-justifying goal, and here "because it feels good" qualifies, while "because a friend told me to" does not.

This is why, if a friend is eating an exotic dish that we consider horrible, there is little to be done other than asking "does that really taste good to you?", wait for the affirmative nod, and change the subject. Unless, of course, you are willing to try and gradually expose your friend to other dishes and educate his taste buds until what "tastes good" has changed. Such an action rests on admitting that, as long as the dish tastes good, this might be a worthwhile reason to eat it.

It is hard to express this point as anything other than a direct appeal to intuition of the reader. But it will surely be granted that we are likely, under certain circumstances, to accept "it just feels good" to be self-justifying in a sense that we cannot accept "because my friend did". It is hard to explain what it is for something to be "self-justifying[22]", yet the idea makes most intuitive sense when considering cases like

[22] "Self-justifying" here means "self-justifying" *to the person* having the experience. It does not imply that an action that is self-justifying in this sense should also be approved of or socially condoned. There need not be any inconsistency between holding that an action is self-justifying *for someone*, while also strenuously opposing it. This point will hopefully be clarified when it is explained why "good" is a phenomenal quality (below) that implies ethical egoism (chapter 2).

"feels good". That "feels good" can be self-justifying means that it is related to our conception of "the Good" by more than linguistic curiosity.

In this thesis it will be argued that our conception of "the Good" ultimately derives from "feels good". Before the argument for such a contention can continue, however, the interpretation of the above "appeal to intuition" needs to be defended against an obvious objection.

3.5. Argument 2: Ethical Hedonism and Psychological Hedonism

It might well be agreed that we do allow "pleasure" to be a *terminus* of any inquiry into goals, without it being conceded that this implies ethical hedonism. A psychological hedonist (such as Schlick) might say that we allow an inquiry into goals to end at this point, but that this is simply because people *want* pleasure. This does not, in any sense, establish ethical hedonism, i.e. that pleasure is what *should* be wanted. Rather it is the result of Schlick's "law of human motivation"; pleasure is simply wanted, and this is a brute fact about people. When "feeling good"/pleasure is allowed to go unchallenged as a goal of human action this is not because it is self-justifying. Rather we are implicitly admitting that "wanting pleasure" is a type of fundamental datum.

One need not be a psychological hedonist to develop this type of objection against the argument advanced above. One might well take the "common sense"-view that people simply want lots of things, and that pleasure is merely one of the many wanted "objects". There is no reason to suppose that these wants are reducible to one primary want, nor that there is any specific justification as to why one thing is wanted, rather than another. This type of view is necessarily absurd if psychological hedonism can be shown to be absurd, as will be explained below.

An ethical hedonist would say that pleasure is wanted because it is "good"; here "wanting" is a secondary phenomenon derived from the nature of pleasure. A psychological hedonist would say that we simply want pleasure, and that we happen to then call "good" those actions that satisfy this fundamental want[23]. Hence an ethical

[23] Hobbes, although not quite a hedonist, held a similar view: "But whatsoever is the object of any man's Apppetite or Desire, that is it, which he for his part calleth Good" (1973: 24).

hedonist would state that we "want what is good", a psychological hedonist simply that we "want what we want".

This type of debate is an old one with different versions having cropped up in a variety of fields. The view that "wanting" is a brute fact, a fundamental *datum*, is related to philosophical debates about the primacy of the will, psychological ideas about a hierarchy of needs, and economic ideas about defining economic actors in terms of a given ranking of preferences[24]. Is it possible to adjudicate between these two claims?

The approach taken here will be as follows. If it is conceded that questions regarding goals will sooner or later encounter a *terminus*, this can only be interpreted in one of two ways. Either the fact that it represents a *terminus* has to do with the goal itself, or it is a brute fact about the person concerned. It will be argued that assuming it to be a brute fact, i.e. assuming "wanting" to be an irreducible primary, leads to absurd implications. Hence the fact of "wanting" does not make sense unless the reason for wanting is located in the object of the want, i.e. unless this object is, in some sense, "good".

If "wanting pleasure" is an irreducible primary, then the notion that human beings want pleasure is, in principle, a contingent fact about human beings. This implies that it is not, in principle, absurd to suppose that something else is wanted. Simply put, if "pleasure is wanted" is an irreducible primary, then this implies that "wantedness" is in no sense the result of anything peculiar to the object of the "want" (pleasure). And if there is nothing peculiar - or "queer" in Mackie's sense - about pleasure that enables it to be wanted, then it is not absurd to suppose that something else could have been wanted.

Consider a being that does not, in any sense, want "pleasure". Imagine that there is a being that is perfectly indifferent between pleasure and pain, and that feelings of pleasure and pain do not ever factor into his decision-making processes as criteria of

---

[24] An interesting version of this debate originates in Plato's *Euthyphro*. Does God approve of what is Good, or does Good simply mean "what God approves of"? See Schlick (1962: 10-11), Mackie (1977: 229-232). Wittgenstein's comments on Schlick's position are in Waismann (1979: 115).

choice. Let the being have a "want" as irreducible primary, imagine that the being wishes to, say, build houses.

Now, if such a being succeeds in building houses, it is difficult to see in what sense this can really matter. If feelings of "satisfaction" or "pride" or "taking pleasure in your achievement" are of no consequence, wherein lies the point of success? If the person fails in building a house, but still there are no "unpleasant experiences" such as "frustration", "disappointment", "anger", etc., then it is hard to see why the person should care. Does the very importance of the difference between success and failure not disappear altogether? Can we still feel sympathy for the failures of the person if any "emotional suffering" becomes irrelevant? There is something intuitively wrong with this picture; it is hard to consider such a "person", driven to build houses, much different from any object under the control of physical law.

The bare fact that a wanted goal was not achieved is not quite enough to arouse our sympathies. We also want to know what happens when such a goal is not achieved. And here the answer "I suffer" or "I feel pain" is sufficient in a sense that we cannot imagine "I didn't build a house" to be. "I suffer" has "to-be-avoidedness" in a way that "I didn't build a house" cannot.

The above "argument" is, in fact, little more than an appeal to intuition. It tries to convince that the assumption of "wanting" as an irreducible primary leads to situations where some indefinable element is clearly missing. This element, then, would be that indefinable element that we refer to as good in-itself, self-justifying or to-be-done. Adding examples that essentially trade on the same point would be pointless; one more will hopefully suffice to show that assuming "wanting" to be an irreducible primary leads to absurd implications.

3.6. Argument 3: Ethical Hedonism and Psychological Hedonism

If there is no *a priori* reason why one object is wanted rather than another, then this equally well applies to what we normally call "pleasure" and "pain". Imagine that pleasure and pain are now reintroduced. Imagine a being that can "feel pleasure" and "feel pain", in the same sense as we normally use these terms, but that the creature

simply wants pain, and does not want pleasure, and that this is an irreducible primary. Imagine a creature that, subject to constraints of knowledge, seeks to live as much of his/her life in agonising pain, and that this is a goal in-itself, a "want" as irreducible primary.

If wanting pleasure or wanting pain was justified only with reference to "wanting", and not with regards to "pleasure" or "pain" as objects of the want, then we could have no absolute preference as to whether we would rather be pleasure–seeking or pain-seeking beings. If there was nothing about pleasure or pain that *causes* us to want the one and not the other, then we might as well have been pain-seeking beings.

But it must surely be conceded that it is absurd to suppose that we can be indifferent between being "pleasure-seeking" or "pain-seeking" beings. We balk at the idea of being forced to seek pain, for is a creature who is doomed to seek unbearable suffering not the most wretched creature imaginable? Is the very possibility of seeking pain for its own sake not unintelligible?

Intuition seems to insist that there is something specific about pleasure that enables it to be wanted, and also something specific about pain that causes it to be avoided. This intuition is one that cannot be explained by any doctrine that takes "wanting" to be a brute fact, i.e. psychological hedonism. Ethical hedonism suffers no similar problem. It states that we cannot be indifferent between pleasure and pain, because pleasure is "good" and pain is "bad"[25].

3.7. Interpretation and Importance of the Three Arguments

Hence it is concluded that "wanting pleasure" cannot be an irreducible primary, since this implies that there is nothing "special" about pleasure that enables it to be wanted. And if there is nothing special about pleasure, then anything else could, in principle, have been wanted, which is absurd.

---

[25] Questions that arise due to the existence of masochism are briefly dealt with in chapter 6, note 150.

The above sequence of three arguments, if accepted, enables a number of related ideas to be rejected. If psychological hedonism fails, then all theories that locate the genesis of action in a "want" or many "wants" *as irreducible primaries* fail in similar fashion. For presumably those that contend that we want many things will also include pleasure among these many things that are wanted. But if we cannot imagine a being that is indifferent between pleasure and pain (argument 2) and cannot imagine a being that wants pain instead of pleasure (argument 3), this implies that there is something "special" about pleasure that enables it to be wanted. And if there is something "special" about pleasure, then this implies that pleasure cannot merely be one wanted object among many wanted objects. Hence the idea of many independent and irreducible "wants", with none being more fundamental than any other, fails.

Non-cognitive "attitudinist" or "emotive" theories of ethics fail in a similar manner. For if "expressing approval" of an action is really all there is to value-talk, then approving or not-approving must be a brute fact or irreducible primary. This is akin to saying that some situations are "wanted" and other "not wanted", i.e. the view that was dismissed above. As soon as there is a "reason", in the sense of a self-sufficient goal, for approving of a state of affairs, then approving can be based on the identification of this fact, and hence cognitive.

In declaring "wanting pleasure" to be a fundamental datum, psychological hedonism dismisses the very possibility of explaining *why* pleasure is wanted. Therefore it cannot explain why assuming other objects to be wanted leads to implications that offend against intuition. Ethical hedonism suffers from no such defect. It states that pleasure is "good" and "self-justifying", and that this causes it to be wanted. Anyone who attempts to ask why the fact that it is "good" is a sufficient reason, simply fails to understand the sense of "good" being used, as explained in argument 1.

It is the main contention of this thesis that this seemingly strange "goodness" or "badness", inevitably encountered when we imagine pleasure or pain, constitutes a legitimate *referential* use of the term "good". In other words, it forms a part of reality, irreducible to anything else, that can only be referred to by value-terms like "good". Hence it qualifies to be Mackie's "metaphysically peculiar" element, the one that is needed to make ethics cognitive.

It remains to give a characterisation of the "good", the "metaphysically peculiar" quality upon which this chapter depends.

5. "Good" as a Phenomenal Quality

5.1. "Phenomenal Qualities"

Below it will be argued that "good" is a non-definable, phenomenal quality (*quale*) that we often refer to as "pleasure" or "consciousness of pleasure". A classic explanation of a non-definable, phenomenal quality is given in Moore's polemic against Naturalistic ethics:

> Consider yellow, for example. We may try to define it, by describing its physical equivalent; we may state what kind of light-vibrations must stimulate the normal eye, in order that we may perceive it. But a moment's reflection is sufficient to shew that those light-vibrations are not themselves what we mean by yellow. *They* are not what we perceive. Indeed we should never have been able to discover their existence, unless we had first been struck by the patent difference of quality between the different colours. The most we can be entitled to say of those vibrations is that they are what corresponds in space to the yellow which we actually perceive (1969: 10).

Phenomenal qualities, or *qualia*, are those qualities that have, to use Searle's phrase, "a subjective, first-person ontology". Simply put, the term phenomenal quality refers to "what colour looks like", "what music sounds like", "what pain feels like", etc. They are those qualities that, even if a complete physical description (size, mass, motion, etc.) of the universe is given, have not been cited. Phenomenal qualities are controversial, some ("qualiaphobes") deny their existence altogether. Others view their existence as self-evident, and use it as an argumentative base from which to establish the irreducibility of consciousness to physiological-neurological states[26].

[26] Most famously Nagel's "What is it like to be a bat?" (1974), which argues that there must be something that it is *like* to be a bat, and that this quality is irreducibly subjective. Also see Searle (1992: 111-126) and Searle (1998: 55-57).

The characterisation of "good" here developed depends on the existence, in some sense, of phenomenal qualities[27]. It is outside the scope of this thesis to present an elaborate argument for their existence. Personally I am convinced by the fact that a being can be without the sense of hearing, and yet give a "full" account of sound in terms of wavelengths, etc. Yet it seems undeniable that a being that can hear knows something about sound that a being without hearing does not. Simply put, the average human being knows what sound "sounds like", the object of this knowledge (however understood) is the phenomenal quality[28].

5.2. "Good" as "Phenomenal Quality"

Schlick states that "...every content of our consciousness, as we learn from experience, possesses a certain *tone*" (1962: 37); in a similar vein Searle asserts that for every "...chunk [of consciousness], it seems to me there is always a dimension of pleasure or unpleasure" (1992: 141). To ignore this results in a "...strikingly joyless picture of pleasure or happiness" , one, which "... seems to lose all the sparkle which life has at its best" (Sprigge, 1988: 132).

Above it was argued that this "tone", (or "pleasure", or "joy" or "sparkle") is self-justifying, or includes to-be-doneness in a sense that other justifications do not. And, since they allow the infinite regress of justification to end in a manner that is not arbitrary, but one determined by the nature of these experiences, they give content to our notion of "good".

A physical (neurological-physiological) account of pleasure can be given, such an account can be complete and yet the term "good" will never be employed. However, it was argued above that the experience of "pleasure" does have something intuitively "good" about it, and that this "goodness" is needed to render our attitudes toward pleasure and pain intelligible. Since this "goodness" does appear to be part of our experience of the world, it is here contended that it is a phenomenal quality.

---

[27] What type of ontology is implied in taking phenomenal qualities seriously is a matter of debate. It would fit with substance-dualism, or (the more respectable) property-dualism. Searle (1992) defends the idea that taking phenomenal qualities to be real does not entail a rejection of physicalism at all.

Phenomenal qualities are intrinsically "indefinable" in Moore's (and Schlick's) sense of the term. But, as stated by Schlick, "we can... make clear what is meant by appropriate indications" (1962: 37). In this manner we can make clear what is meant by "yellow" by saying that "yellow" refers to what it *looks like* when light-vibrations of a certain kind stimulates the eye. In a similar manner it can be made clear what is meant by "good" by saying that "good" is what it *feels like* when we are in a certain physiological-neurological state, one that we often call "pleasant"[29].

It is important to be very clear about the claim made above, as will be shown when Moore's "open question"-argument is discussed in chapter 2. It is being maintained that pleasure (phenomenal) is *identical* to what should properly be characterised as "good", *not* that "goodness" can be predicated of pleasure (phenomenal). In other words, it is being argued that pleasure (phenomenal) and "good" refer to the same object, and that it has not commonly been realised that this object is metaphysically strange in that it can end the regress of justification in a non-arbitrary manner. Hence "pleasure" (phenomenal) *can* be used to refer to "good", but then it functions as an irreducible value-term.

It is *not* being claimed that pleasure is "valuable", but simply that it *is* "value". Saying "pleasure (phenomenal) is valuable" is akin to saying that "colour is coloured" or "length is long". Rather we say that, for instance, "the table is coloured" or "the stick is long". In a similar way it can be said that a certain state "is valuable", if this is taken to mean the same as saying that "pleasure" (phenomenal) or "goodness" is an attribute of such a state.

It remains to clear up a terminological issue.

---

[28] This is a version of Jackson's "knowledge argument". For a discussion, a more careful formulation of phenomenal qualities than the Moorean one cited above, and a discussion of the argument between "qualiaphobes" and "qualia-freaks" in general, see Burwood (116-137).

[29] Treating "good" as a quality of a subjective, first person state means that it is impossible for a third party to have direct access to facts concerning "good". This is a consequence of taking first person states seriously. Viewing the "good" as a *quale* implies that they can be directly known only to the person having the experience, and that the first person only knows them by "acquaintance". My choice for the first person view was explained in the introduction, and will not be defended here.

6. The "Value of Experience"

"Pleasure" (phenomenal) is often used without the implication that it is an irreducible value-term being realised. This is much less likely with "feeling good", therefore "feeling good" is a better characterisation than "pleasure". For this reason the word "pleasure" will be discarded wherever possible.

There are also less important reasons why dropping "pleasure" seems a good idea. "Pleasure", because of its most common usage to refer to "bodily pleasure", also doesn't always easily communicate all that can be communicated by "feeling good". "Pleasure" seems somehow inadequate to capture feelings of ecstasy or triumph, "feeling good" (or "feeling great!") works better. The phrase "intellectual pleasures" also seems oddly unsatisfying (to me, at least). Experiences like, for example, solving a difficult problem, or having a previously impenetrable argument suddenly become clear in a moment of insight, can be among the most profound imaginable. To describe these as "extremely pleasurable" or even "enjoyable" almost seems to belittle them. On the other hand, if you have just solved Fermat's Last Theorem and say that you simply feel "indescribably good", you are less likely to feel that you are insulting the experience.

The meanings of "pleasure" and "pain" seem to need some uncomfortable stretching to include all that can be covered under "feeling good" or "feeling bad". And yet it is not immediately evident what could adequately replace them. "Hedonic tone" is too closely linked to "pleasure" and "pain". Meinong's phrase "value-experience" remains peculiar to himself and is conceptualised in a way radically different from the conceptualisation of "feeling good" in this chapter.

The phrase "value of experience" (where value means roughly "worth") seems most suited to convey the sense of "goodness/badness of experience", and will be used for the rest of this thesis. "Value", as explained at the start of this chapter, will be used as a technical term to denote "that aspect of experience that can only be expressed using irreducible value-terms". The main claim of this chapter then becomes that the "value of experience" is a phenomenal quality of a physiological-neurological state that has the special property of ending the regress of justification in a non-arbitrary manner.

(The term "value" also has other meanings (numerical value, etc.) that will be employed in this thesis, but the context should be clear enough to prevent any confusion.)

7. Conclusion

This chapter firstly attempted to determine what would be needed for value-judgements to be cognitive, and hence a delineable *subspecies* of factual judgements. The requirement of cognitivity was shown to amount to the requirement that a legitimate referential use of "good", as irreducible value-term, be identified.

This requirement can also be stated as the requirement of a non-arbitrary end to the regress of justification. It was then argued that what is commonly called "pleasure" is often allowed to end the regress of justification. It was also argued that considering "pleasure" as justification gives us the clearest intuitive sense of what it is for this regress to end in a *non-arbitrary* (and self-justifying) way.

Attempts to explain our attitudes towards pleasure by making the idea of a "want" an irreducible primary (as is done by psychological hedonists) were dismissed as incoherent. Our attitudes toward pleasure only gain intelligibility if pleasure is identified with "good", i.e. allowed to be self-justifying by its very nature. This leads to the theory of cognitive, ethical hedonism. The main novelty of this position, as developed so far, is that it is cognitive, i.e. it allows for value-judgements to be on par with factual judgements.

The object of this referential use of "good" was characterised as a phenomenal quality. It can be defined as "what pleasure (physiological-neurological) feels like". This phenomenal quality will hereafter be referred to as "the value of experience".

The position outlined above is cognitivist and hedonist. Numerous arguments have, over time, accumulated against these positions. This is, indeed, to be expected of positions that have been around for ages. The next chapter will attempt to refute the most important or obvious of these arguments.

Chapter 2: Cognitive, Egoist, Ethical Hedonism

1. Introduction

The conceptualisation of "good" developed in chapter 1 leads to a theory of cognitive, egoist, ethical hedonism. There are various conceptual difficulties with, and well-known objections to, the various components of such a theory. This chapter will attempt to clarify the most important of these difficulties, and to answer the most serious of these objections.

The first component that will be considered is "hedonism".

2. Cognitive, Egoist, Ethical *Hedonism*

2.1. Hedonism – Characterisation and Clarifications

In chapter 1 it was argued that what we commonly refer to as "pleasure" is more adequately characterised as the irreducible value (goodness/badness) of experience. Hence the argument for hedonism has, too a large degree, already been made. A few points still stand in need of some clarification.

The first concerns the conceptualisation of "good" as a phenomenal quality. Phenomenal qualities, by definition, relate to the experience of the world, and not the world itself. This poses the question of the relation between phenomenal qualities and the world, or between phenomenal and physical facts. It will be assumed that changes in phenomenal facts can mostly be correlated to changes in physical facts. For example, if I add wood to a fire and it grows bigger (physical facts), then I will feel warmer (phenomenal fact). This appears to be no more than mere common sense, but is important since it implies that an explanation of a change in phenomenal states can be given by referring to a change in correlated physical states[30].

[30] While it will be assumed that changes in physical and phenomenal states are correlated enough for this type of explanation to be possible, it need not be assumed that they are always, and precisely, so correlated. The idea of a strict supervenience between phenomenal and physical states might well be the most elegant and intuitively sensible way to conceive of this relation, but nothing in this thesis depends on the defence of such a claim.

Another important implication of the characterisation of "good" as a phenomenal quality is that "goodness", in the literal sense, can *only* exist where there is experience. Of course, there are a multitude of other things, besides experience, that are often referred to as "good". These can include actions, paintings, personality-traits, moral codes and so on, *ad infinitum*. While it will be maintained that none of these things can, literally and by themselves, be good, these uses of "good" will not be dismissed as unimportant. Rather these uses might well be explained as ultimately deriving from the value of experience. This explanation will be attempted from chapter 3 onwards.

It will be assumed that all experience has value, i.e. that it can be asked of any experience whether it is good, bad, or, metaphorically speaking "in between"[31]. If this is not the case, i.e. if the question regarding the value of experience is sometimes simply inapplicable, this would not seriously matter. Such cases would simply disappear from consideration.

It will be assumed that the difference between "good" and "bad" experiences can be spoken of as a matter of degree. In other words, any two experiences must either have the same value, or stand in a relation of "better" or "worse". This is another way of stating that there exists an ordinal ranking of all experiences that is both complete and transitive. (Whether this ranking can also be thought of as cardinal is discussed below.)

A final word about "incorrigibility". Is it possible to be wrong about the value of your experience, while having the experience? A number of philosophers who have no quarrel with the idea of subjective, qualitative experience have denied that any reports of such experiences are incorrigible[32]. Nothing much in this thesis seems to rest on deciding this issue either way. Hence issues regarding "the incorrigibility of pain-experiences" will be ignored.

---

[31] Compare Searle (1992: 141): "For such a chunk [of consciousness], it seems to me there is always a dimension of pleasure and unpleasure. One can always at least ask some questions in the inventory that includes "Was it fun or not?" "Did you enjoy it or not?"..."

2.2. Objections Related to "Philosophical Grammar"

One oft-cited objection to hedonism is that it misunderstands the grammar of "pleasure" and "pain". It is sometimes claimed that "pleasure" does not directly refer to a set of experiences that have certain attributes in common. Rather "pleasure" is a type of umbrella-term for a number of experiences that are, at best, loosely linked in a manner akin to Wittgensteinian "family-resemblances".

If all uses of "pleasure" are taken into account then something like the above is probably true. This difficulty is removed by the narrower conception of the "value of experience" developed in chapter 1. The value of experience is supposed to refer to that aspect of experience that cannot be expressed except by employing irreducible value-terms. Any uses of "pleasure" that do not refer to such an object are irrelevant to this thesis, and excluded from consideration[33]. (It might, of course, be doubted if anything satisfies this requirement. This is a different issue, and was discussed in chapter 1.)

There are other possible objections based on the grammar of "pleasure" and "pain"; these are mostly due to the behaviourist strains in the later Wittgenstein, Ryle, etc. For example, a behaviourist might well object to the conception of "pleasure" as inner experience. Rather a behaviourist might define "pleasurable action" as "action we are likely to persist in". This would render any claim that people seek pleasure tautological.

The above objections to the uses of "pleasure" or "pain" normally form part of a more general scepticism regarding the "folk psychology" behind intentional and teleological descriptions as such. The issues involved in such arguments tend to hinge on complicated questions regarding the philosophy of language and the conception of mind. To truly take all such considerations into account would require an enormous amount of time, and for this thesis to run into several volumes. The situation is further

---

[32] For example, see Searle (1992: 144-149).

[33] Another possibility would be to admit that certain parts of our experience are inexpressible except through value-terms, but to doubt that this is simply, at any given time, a brute fact. Rather the value of experience can be conceived as a composite of co-existent value-components. It is being assumed that

complicated by the fact that several of the questions concerning language and mind might well depend our understanding of value-judgements. Hence I will not do much more than to once again cite Neurath's oft-repeated metaphor: if a ship is to be rebuilt at sea, then we need somewhere to stand. In this case the "somewhere to stand" amounts to an acceptance of intentional and teleological description, i.e. that experience is irreducible to action (or "outward criteria") and that people can be said to "strive after" something.

The next issue to be discussed concerns "cognitivity".

3. *Cognitive*, Egoist, Ethical Hedonism

3.1. Cognitivity – Definition and Problems

In chapter 1 the idea of "cognitivity" was defined in terms of facts and truth. There are many different accounts of fact and truth, and there seems no particular reason to suppose that the main ideas of this thesis are compatible with only one of them. A minimum proviso, stated in chapter 1, is that truth not simply be *identified* with value. If this type of crude pragmatism is dismissed, then fact and value are conceptually distinct. This raises the question of the relation between them. The central aim of this thesis is to portray this relation as that between a class and sub-class, i.e. to present a theory on which value-judgements are a delineable *sub-species* of factual judgements. Or, in other words, "ought" becomes a type of "is".

There are certain well-documented difficulties with such an attempt. The most general concerns the so-called "Naturalistic fallacy", both in its Humean and Moorean form. This will be the first problem considered below. The other concerns the question in what sense value (as defined) is the type of thing that can be said to be capable of being *maximised*. This is an extremely difficult problem, and will receive substantial treatment below.

---

is not the case, but if this assumption were false it would not matter a great deal. Rather the problem of aggregation, discussed below, would rear its head one step earlier.

3.2. The Naturalistic Fallacy – Hume

The term "naturalistic fallacy" was originally used by G E Moore to refer to a mistake he thought typical of ethical reasoning. The essence of the problem was originally pointed out by Hume. His version is also known as "Hume's law" or the "is/ought"-fallacy.

In Hume's *Treatise* he states that writers commonly argue about "what is", and "is not", and then suddenly proceed to use the terms "ought" or "ought not". This is surprising, because " …as this *ought*, or *ought not*, expresses some new affirmation or relation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it (1969: 521). An example of such a doctrine would be crude social Darwinism. It might well argue that society is the result of survival of the fittest. If it tries to derive the conclusion that the fittest *should* survive from this "data", it violates Hume's law.

Simply put, Hume is saying that no statement that includes an ought can be logically implied by a statement that relates to what is. Such a relation of logical implication is only possible if there is a premise connecting the two, but writers do not make this claim. Rather they just suddenly make the leap from "is" to "ought". He adds that any claim regarding a relation between "is" and "ought" that would justify such inference seems "altogether inconceivable", for they are "entirely different".

On this formulation it must surely be apparent that Hume's problem does not add anything to the problem identified when discussing the *Tractatus* and Mackie's "argument from queerness". If the statement "You should do what is good" is taken to be a tautology, then saying that "is" and "ought" are "entirely different" amounts to the same as saying that ethics has no "object". In chapter 1 it was argued that ethics does have an object, i.e. that "goodness" or "to-be-doneness" is instantiated. If this is correct then value-judgements form a subset of factual judgements in the same sense that judgements about any particular thing form a subset of factual judgements as such. Hence there is no particular problem about the cognitive status of value-judgements as such.

The Humean version of the Naturalistic fallacy does not present any *additional* problem for the idea of a cognitive ethics; rather it amounts to a reformulation of the point under dispute in chapter 1. The same can, in a sense, be said about Moore's version of the argument. Still it will be useful to discuss as it allows for the clarification of some related issues.

3.3. The Naturalistic Fallacy – Moore

3.3.1. Clarifications

Two points need to be clarified before Moore's argument is discussed. Firstly, he defines hedonism as the view that "pleasure is the sole good" (1968: 69). This is not the contention being defended in this thesis, rather it is being claimed that what we often call "pleasure" is identical with "good". This was explained in chapter 1. Any objection of Moore's (to the view defended in this thesis) that relies on this definition will not be considered.

The second concerns the main aims of *Principia Ethica*. A large part of Moore's writing is directed at a defense of the claim that "good" is "simple" or indefinable. This is to say that it is not a "whole" made up of "parts" that can be used for a definition, in the sense that a "horse" can be defined by its parts as an animal "…[having] four legs, a head, a heart, a liver, etc." (1968: 8). The conceptualisation of "good" developed in this thesis does not dispute Moore's claim that "good" is indefinable in this sense. In chapter 1 it was, in accordance with Schlick, agreed that all phenomenal qualities are indefinable, and that we can "only make clear what is meant by appropriate indications"(1962: 37). The thrust of Moore's argument is against "naturalistic ethics", i.e. any attempt to define good in terms of natural qualities. Since this thesis does not, in Moore's sense, attempt to define "good", it does not really constitute a naturalistic ethics.

3.3.2. The "Open Question-Argument": Statement

Moore would, however, still reject the claims made in this thesis. For, although "good" is not being *defined* in terms of "pleasure", it is being asserted to be identical

to what is often called "pleasure". Rather Moore would claim that "pleasure" and "good" are distinct, so that the statement "pleasure is good" is true and synthetic (1968: 9). Pleasure, on Moore's account, is simply *one of* the good things that are to be found in the world (1968: 9).

Moore's reasoning in dismissing any identification of pleasure with good rests on his celebrated "open question-argument". Consider the question: is pleasure good? This question can only make sense if "pleasure" and "good" are distinct. Hence, if it is admitted that the question makes sense, then it is thereby admitted that "good" is not identical with "pleasure" (1968: 9-12).

3.3.3. The "Open Question – Argument": Identical

The first consideration when responding to Moore concerns the idea of pronouncing one thing to be identical with another. There is a *prima facie* sense in which any such undertaking is absurd, for one thing cannot, by definition, be another. This truism obviously inspired much of Moore's thinking in his *Principia*; the title page contains Bishop Butler's statement that "[e]verything is what it is, and not another thing". Hence it needs to be clarified in what sense it is not absurd to say that one thing is another.

The above difficulty is solved by realising that the same object can be identified by different criteria. Frege's famous example in drawing the distinction between sense and reference illustrates this point well. The criteria whereby "morning star" is identified differ from the criteria whereby "Evening star" is identified, yet they both refer to what is also called "Venus". "Evening star" and "Morning star" are co-referential. Hence I can say "the Evening star is the Morning star", and, if I simply mean they are co-referential, this is correct. This is one sense in which one thing can be said to be identical with another, where "identical" simply means "co-referential". This is equivalent to what is being claimed in this thesis. It is being claimed that pleasure (phenomenal) is co-referential with "good" (or "to-be-doneness" or "end-the-regress-of-justification-in-a-non-arbitrary-mannerness", etc).

3.3.4. The "Open Question – Argument": Pleasure is Good

The above demonstrates one possible sense in which a claim of "identity" can be made sense of. The case with "pleasure" and "good" has a difficulty that does not occur with "Evening star" and "Morning star". This is that both "Evening star" and "Morning star" are complex objects, in Moore's terms, i.e. they are both definable. This is not the case with "pleasure" and "good". Moore contends that both are indefinable, simple[34] objects. Hence the difficulty boils down to this: Moore finds two simple, indefinable objects, whereas I can find only one.

In chapter 1 it was argued that only the consideration of what is often called "pleasure" gives a type of intelligibility to the idea of something being "self-justifying". While this would be consistent with the idea that "pleasure is the sole good thing", such an interpretation seems needlessly profligate. Rather an application of Occam's razor leads to the conclusion that pleasure simply *is* goodness-itself.

Moore, however, thinks there is a need to postulate an extra entity. He states that, if anyone asks "Is pleasure good?", he quickly realises that, concerning "goodness", he "has before his mind a unique object" (1968: 16). Is he correct? It will not be disputed that there appears to be a distinct entity before the mind when such a question is asked. Rather, and in the spirit of Occam, it will be argued that this "appearance" can be accounted for without reifying this entity.

It will surely be conceded that words like "pleasure" can be used in a variety of senses. Consider "pleasure (phenomenal)", "pleasure (physiological-neurological)", "long-term pleasure", "immediate pleasure", etc. Hence it is possible that, if someone claims to have in mind a clear conception of the good, distinct from "pleasure", the sense of pleasure used is not the one proposed in this thesis as the "object" of ethics. In fact, it can be argued that the mind has a habit of interpreting statements in such a manner as to give them a sense. This habit then leads to assuming a sense of "pleasure" that makes the statement "pleasure is good" concern two different objects.

---

[34] Moore is in agreement with Schlick's assertion that "pleasure" is indefinable (1968: 13).

It will be argued in chapter 5 that, while pleasure should be maximised, this is, in-itself, a poor guide to action. In other words, pleasure can be better maximised by consistently trying to achieve a set of distinct objectives, which are much easier to follow than "maximise pleasure". This would then lead to the possibility that, when we ask of something whether it is good, the "objects before our minds" concern these distinct objectives. Such a situation would clearly give illusory support to Moore's point of view, but explain why it seems to make sense.

There is another powerful reason for not trusting Moore's assertion that, when considering goodness, he has before his mind a distinct object. This is based on the insight for which Wittgenstein is famous. If Wittgenstein is correct in saying that language can mislead in such a way that nonsensical statements are mistaken for "real" propositions, then this leads to the possibility that the sense of "good" before the mind of Moore is a linguistic illusion[35]. If there is merit in Wittgenstein's claim that language can give rise to such confusions, then Moore's conviction that he has a clear sense of "the good" before his mind is not to be taken at face-value.

Hence it has been shown that Moore's argument cannot function as a knock-down argument against the type of hedonism being defended in this thesis. Furthermore, some of his central insights are even preserved in this thesis. But a combination of these insights with Occam's razor and the arguments presented in chapter 1 leads to a dismissal of any argument he might have against the type of hedonism argued for in this thesis.

There is another, more subtle point that needs to be highlighted before proceeding. On page 60 of *Principia Ethica*, Moore writes:

> In ordinary speech, 'I want this', 'I like this', I care about this', are constantly used as equivalents for 'I think this is good'. And in this way it

[35] Austin has claimed Moore's conception of "good" to be such a linguistic confusion: "If someone did not know about cricket and were obsessed with the use of such 'normal' words as 'yellow', he might gaze at the ball, the bat, the building, the weather, trying to detect the 'common quality', which (he assumes) is attributed to these things by the prefix "cricket". But no such quality meets his eye; and so perhaps he concludes that "cricket" must designate a non-natural quality, a quality not to be detected in any ordinary way but by intuition. If this story strikes you as too absurd, remember what philosophers have said about the word 'good' …" (1962: 64).

> is very natural to be led to suppose that there is no distinct class of ethical judgements, but only the class of 'things enjoyed'...

The above reasoning is not totally dissimilar to the reasoning employed in chapter 1, but the conclusion drawn was the direct opposite of that reached in the above passage. Moore is against the reduction of the predicate "good" to a natural quality. This is reasoning that is commonly found in a psychological hedonist position, and can rightfully be charged with having committed the naturalistic fallacy. But that it is not the case with the argument advanced in this thesis. Here "good" is not being reduced to "pleasure", rather "pleasure" is being, metaphorically speaking, "upgraded" to the status of "good". Hence the idea of value in the theory of ethical hedonism is not being used to dismiss the idea of a distinct class of ethical statements. Rather it is being argued that certain statements about "pleasure", fully understood, *constitute* the distinct class of ethical statements.

It has now been shown why the argument in the first chapter, if accepted, dissolves any problems vis-à-vis the naturalistic fallacy, and some related issues have been clarified. There is another problem with the idea of a cognitive ethics that specifically relates to hedonism. This is the problem concerning the meaning of *maximisation*, and needs to be discussed at length.

3.4. Cognitivity – Maximisation

3.4.1 Statement of the Problem

There exists a problem regarding the ordinality or cardinality of pleasure, and hence "goodness". A scale is ordinal if it is simply a ranking of certain things. For example, the horses in a horse-race can be ranked as first, second, third, fourth, etc. A scale is cardinal if it ranks certain things relatively to each other, but also defines the position of these things with reference to more than their relative position. Here they also have an "absolute" position according to a given scale. Take a ranking of horses in a race that orders them according to the time it took them to complete the race. Here each

horse has a relative position, but also an absolute position in terms of the time it took to complete the race.

An ordinal scale only gives information regarding relative position, and cannot answer questions like "was the difference between first and second greater than the difference between second and third?". A cardinal scale is an absolute measure, and can answer questions like the above and questions regarding relative position. Note that any cardinal scale also implies an ordinal scale. In other words the cardinal measure "time it took to past the post" implies the possibility of an ordinal measure "position when passing the post".

Is "pleasure" or "goodness" an ordinal or cardinal measure? In other words, can it only be used to rank given alternatives as more or less pleasurable, or can it assign some absolute magnitude to each alternative?

An easy way to distinguish ordinal and cardinal scales is with reference to the difference between the positions of ranked elements. On an ordinal scale this difference has no meaning, it tells us nothing about the relative difference between third and fourth and fourth and fifth. On a cardinal scale this has meaning, each element has a position based on an absolute commensurate scale. This means that mathematically it is possible to perform certain mathematical operations, for instance adding and subtracting, on cardinal numbers. For instance I can subtract the time it took A to complete the race from the time it took B, in order to get the *difference* between their times. This would be nonsensical with ordinals.

The above forms the main reason why the issue of cardinality is relevant to this thesis. This thesis defends the view that the subject should maximise value. This amounts to the view that the option that will yield the most pleasure over time should be chosen. For the idea of "most pleasure over time" to make sense, it needs to be possible to add and subtract the different levels of pleasure experienced by the subject at different times. This is only possible if pleasure is cardinal, not if it is only an ordinal measure. Hence it is central to this thesis that pleasure should be a cardinal measure.

3.4.2 Argument for Cardinality

The central difference between ordinality and cardinality is whether the difference between the positions it assigns to things make sense. Consider the relative "goodness" of two moments of experience. On the one hand it is obvious that this goodness can lead to an ordinal measure; one will be preferred over the other. But the preferred experience could have been even better. Simply put, however good the experience was, it could have been slightly better or worse. This can be the case *without influencing the ordinal scale.* In other words it makes sense to say that one option can be better than another, but another can be *much better* than the first. If these expressions make sense then pleasure does not only produce an ordinal scale. It also gives the experience a position according to some other measure, independent of the other experiences it is ranked against[36]. In other words, pleasure is cardinal.

The central fact in the above argument is that anything that is good, can be better, without this necessarily meaning it changes position on the ordinal ranking[37]. Hence pleasure reveals more than just an ordinal ranking, it also reveals the existence of an independent and absolute scale.

If pleasure is cardinal, then the phrase "maximisation of value" has a determinate sense. The next issue that stands in need of clarification is that of egoism.

---

[36] The above argument to the effect that there exists a cardinal scale, should not be taken to imply that people's knowledge as to this cardinal scale is always correct, nor that they can always remember their previous rankings correctly. It just implies that there is a consistent scale that can count as a matter of fact in determining whether their judgements are correct or incorrect.

[37] This is not the case with "utility" in "revealed preference theory", as will be seen in chapter 3. Utility is only defined in terms of behaviour, and hence cannot result in cardinal rankings. One exception to this is von Neumann – scales, where behaviour with regards to lotteries is used to construct cardinal scales. Von Neumann cardinality won't do for this thesis, since it does not portray the "behavioural cardinality" as a measure or expression of any "internal cardinality". For this reason it will not be discussed in this thesis. Hence "cardinality" above should be read as referring to only "internal cardinality".

4. Cognitive, *Egoist*, Ethical Hedonism

4.1. Egoism and Phenomenal Qualities

It has already been mentioned that ethical hedonism leads inevitably to Egoism. The characterisation of "good" as a phenomenal quality must surely make it clear why this is the case. The sense in which pleasure can be referred to as "good" is the "phenomenal" sense of pleasure, meaning that it has an *irreducibly subjective ontology*. Hence the sense in which pleasure is a "self-justifying" (ethical hedonism) and motivating (psychological hedonism) goal of action is irrevocably tied to the individual experiencing it. Simply put, I cannot feel the feelings of another, and therefore the "to-be-doneness" of another's pleasure does not "exist" for me.

The assertion that egoism is not a lamentable fact about human nature, but rather represents the "good" in matters of ethics might strike some as strange. Here it is important to distinguish between the ultimate goal of action and the strategy (set of decision-rules) that best accomplishes it. The fact that the ultimate goal of action should be egoist does not, of itself, say anything about the behaviour that is prescribed by such a goal. It will be argued, from chapter 5 onward, that the actual recommendations of an egoist ethics is roughly consistent with our ordinary moral intuitions. The argument is an extremely complex one, but I will at least ask the reader to suspend judgement until the issue is discussed.

Two remaining issues concerning egoism will be dealt with here. The first concerns the logical consistency of egoism, the second the historical association of hedonism with utilitarianism.

4.2. Objection – Logical Consistency of Egoism

Moore has claimed that egoism is logically inconsistent.

> What Egoism holds, therefore, is that *each* man's happiness is the sole good - that a number of different things are *each* of them the only good

> thing that there is - an absolute contradiction! No more complete and thorough refutation of any theory can be desired (Moore, 1968: 99).

Moore is saying that the very idea of egoism is self-contradictory, that I am saying that a class of certain things are said to have a certain property, and yet also saying that each of these things have this property exclusively. Clearly Moore is right in calling this contradictory, three chairs cannot be red and yet one of them be the only red chair. But it will presently be shown that it is Moore's formulation of egoism that is to blame for this contradiction. This is not at all what egoism is saying. For egoism does not hold that each man's happiness is the "sole good". It holds that each man's happiness is the "sole good-for-him"! Phenomenal qualities are, as Searle puts it, *irreducibly subjective*. If "goodness" is conceptualised as a phenomenal quality then the idea of a "sole good", understood as something distinct from sole "good-for-him", becomes a chimera. If Moore's statement is reformulated, substituting "good-for-him" for "good", the contradiction disappears:

> What Egoism holds, therefore, is that *each* man's "good-for-him" is the sole "good-for-him" - that a number of different things are *each* of them the only "good-for-the person concerned".

It is analogous to asserting that something is "painful". Ten different experiences can be "painful" to the specific people concerned, and each specific experience can be the "sole painful" thing to the person concerned, without this implying a contradiction. The class of these experiences can be described as "painful-to-someone", yet each experience can be the "sole painful" experience to a "specific someone".

The possibility that "good" is irreducibly subjective is something that Moore seems to consider, but dismisses.

> It is obvious, if we reflect, that the only thing which can belong to me, which can be *mine*, is something which is good, and not the fact that it is good (Moore, 1968: 99).

Moore is saying that the *thing* which is good can be mine, but that the *goodness* of it cannot. He proceeds to "argue" for this conclusion, but his arguments amount to restatements rather than adding something new to the above statement:

> The *good* of it can in no possible sense be 'private' or belong to me; any more than a thing can *exist* privately or *for* one person only (Moore, 1968: 99).

It has already been shown that phenomenal qualities such as "painful", in the sense of "experiencing pain", is something which, in a loose manner of speaking, only exists-for-me. Something which "hurts-me" can only be described by a third party as "hurting-him", but cannot be described by this same third party as "hurting-me". In this sense, "hurts-me" exists only for the person being hurt. And why does "good" exist like the "redness" of chairs, and not the "hurting-me" of pain? No answer is given by Moore, only reformulations of the assertion that it does not (1968: 98-101)[38].

There is another issue which might well occur to the reader. The traditional formulations of hedonism have predominantly led to philosophers advocating utilitarianism, not egoism. Yet the combination of cognitivity and hedonism cannot imply utilitarianism, as will be shown below.

4.3. Egoism and Utilitarianism

The following seem to be the only two forms in which the argument for Utilitarianism can be put, *while staying within the realm of cognitive ethics*:

1. *Pleasure* is good, therefore more pleasure for more people is better than less pleasure for less people.
2. *My-pleasure* is good, therefore more pleasure for more people is better than less pleasure for less people.

[38] For support of the idea that there is no logical problem with an agent-relative notion of good, see Broome (1994: 129).

The first premise of the first argument, the statement that "pleasure", *understood as distinct* from "my-pleasure", is good, has been argued to be false. This distinction between "pleasure" and "my pleasure" was discussed above, and why it is false was argued above. What makes pleasure "good" is irrevocably tied to a specific person, and then it is good for that person, in the same sense as that which makes pain "painful" is irrevocably tied to a specific person, and then it is painful for that person. If premise 1 of the first argument is amended by substituting "my pleasure" for "pleasure" in order to make it true, and the rest is left unchanged, the argument no longer follows. However, if argument 2 is amended by substituting "pleasure" for "my pleasure" throughout, the conclusion again follows from the premise. This becomes:

3. My-pleasure is good, therefore more "my-pleasure" for more people individually is better for more people individually than less "my pleasure" is for less people individually.

But the above is a formulation, albeit inelegant, of egoism, not of utilitarianism. Thus Utilitarianism must either be false (argument 1) or illogical (statement 2). If the characterisation of pleasure in chapter 1 is accepted, then turning it into a cognitive doctrine implies that utilitarianism becomes egoist, as shown above.

The above can be illustrated by considering Bentham. If he is interpreted as trying to build a cognitive ethics, he can be read as making both argument one and two alternatively:

> Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do (Bentham, 1982: 11).

Bentham is saying that pain and pleasure are the regulators of all human conduct, and the determinants of what should be done[39]. If by this he means "pain-for-me", this is

[39] Problems with regards to whether Bentham's "ethics" can be made consistent with his psychological hedonism will, for the moment, be ignored.

what has been argued to be the case. But, of course, this is not quite what Bentham means:

> By utility is meant that property in any object, whereby it tends to produce...pleasure... or to prevent ...pain... to the party whose interest is concerned: if that party be the community in general, then the happiness of the community: if a particular individual, then the happiness of that individual (Bentham, 1982: 12).

The case of the "community in general" does not make sense. In what sense can the "pleasure-for-me" of a community be increased? This amounts to the formulation "pleasure-for-me-for-the-community", a nonsensical phrase.

Of course, Bentham is not referring to "pleasure-for-me", but "pleasure-for-anyone". But in this sense the term "pleasure" is simply no longer the arbiter of human action. There is a profound difference between "pleasure-for-me" and "pleasure-for-him" that allows the former to be the guidepost for all human action, but not the latter, as argued earlier. So, in this sense, Bentham's original claims regarding "pleasure/pain" are now false, and an instance of the argument previously identified as argument one.

Bentham tries to get around this objection by pretending that the community is a "person", using the Hobbesian notion of the community as a body:

> ... The community is a fictitious *body,* composed of the individual persons who are considered as constituting as it were its *members* (Bentham, 1982: 12).

The metaphor of seeing a society as a "body" surely has its uses, but is only a metaphor, apt in some cases and misleading in others. In the case of "experiencing pleasure", it surely is misleading. The sense in which “pleasure” is good has been argued to be irreducibly subjective. Hence the “experience of the community” can

mean no more than the experience of the individuals composing it[40]. Here exists a disanalogy which disqualifies the use of this metaphor to establish Bentham's principle.

Bentham's argument then boils down to an instance of what was referred to as "argument 1". It rests on the false claim that pleasure, understood as something distinct from "pleasure-for-me", is the arbiter of human action. Or the above can be read as an instance of argument two, since Bentham concedes that individual pleasure can be an ethical guidepost. But he then errs logically when trying to extend this to society by using an inapplicable metaphor.

I would contend that the historical connection between utilitarianism and hedonism has a psychological, rather than a logical cause. If someone is convinced that there is something strange about pleasure that qualifies it to be the sole object of morality, it appears to be mere common sense that no-one's pleasure should count more than another's. This seems to be due to an uncritical application of a principle of equality or fairness that is so basic to our ethical reasoning that it is not commonly realised that it is in need of justification. The acceptance of this principle, combined with hedonism, then leads to utilitarianism. It is unjustified since, what makes pleasure "special", is always relative to the individual experiencing it.

In chapter 6 it will be claimed that one might be able to derive some sort of principle of fairness from egoist principles. The important point, however, is that, if hedonism is to be cognitive, this principle needs to be derived from an egoist base. Jumping from hedonism to utilitarianism in one easy step simply won't do.

With these questions concerning egoism out of the way, it still remains to clarify the exact relation between ethical hedonism and psychological hedonism.

---

[40] Unless, of course, the phrase "communal experience" can be given a quite literal sense. Sprigge manages this by combining hedonism with panpsychism in *The Rational Foundations Of Ethics* (1988).

5. Cognitive, Egoist, *Ethical Hedonism*

5.1. "What Should be Done"

The exact meaning of the ethical "should" has been a matter of frequent dispute in philosophy. Many philosophers have doubted the very intelligibility of a specifically ethical "should". The remarks above concerning the naturalistic fallacy has already shown the essence of why this need not trouble an ethical hedonist. Some remarks concerning the relation between the conditional and absolute "should" should suffice to remove any confusion regarding this matter.

Kant's "categorical imperative" was intended to be a specifically ethical "should", i.e. to be understood as distinct from the unproblematic "conditional" ("hypothetical[41]") "should". Schlick summarised one often-made[42] criticism when he stated that:

> It is just as if Kant had said: "I wish to use the phrase 'to take a walk' with such a meaning that I can say 'a walk is being taken' without anyone there who takes it." (1962: 112).

If the "should" does not command absolutely, but always "hypothetically", then any statement expressing a "should" is relative to the specific goal being expressed. If this "goal" is questioned, then an infinite regress follows where this goal has to be justified with reference to another goal, etc. But if this process is to be stopped an "absolute should" is needed. Yet the very idea of an "absolute should" seems unintelligible.

Ethical hedonism can offer a way out of the above predicament. If some element of experience is found to be good "in-itself", i.e. if a valid referential use of the term "good" can be found, then it can be said that someone should seek this "good". The use of the term "good' here stops the infinite regress. If someone was to ask why (s)he

---

[41] "All imperatives command either *hypothetically* or *categorically*. Hypothetical imperatives... declare a possible action necessary as a means... [a] categorical imperative... as objectively necessary in itself" (Kant, 1964: 82).

should do what is "good", then such a question can only be asked if the person does not understand the meaning of "good" in this context.

The "should" commands "hypothetically", but an infinite regress does not arise. If something is "good", then it is thereby "self-justifying"; no action requires a sanction over and above being "good". An alternative way to phrase this would be to say that "you should do what is good" is an "absolute should", for to ask what further goal is served by actions that are "good" is to misunderstand the meaning of the term "good". The "should" here appears to be "absolute", but only because of the self-justifying nature of "good".

The "value terms" under discussion, terms like "good", "should", "right", form a cluster of mutually defining notions. If one of them gains definition, then all the others are also thereby defined. Hence, if it is allowed that there is a referential use of the term "good", then questions about "should", etc. are also thereby answered.

5.2. "What Is Done"

It was stated earlier that ethical hedonism, i.e. the doctrine that pleasure *should* be sought, implies a form of psychological hedonism, the doctrine that pleasure *is*, in some sense, what is sought. How is this possible?

Imagine a being with perfect information about her environment that has to choose between two options, the one leading to pain and the other to pleasure. Which *should* she choose? She should choose the one that is "good", i.e. the one that leads to pleasure, as was discussed above.

Which *will* she choose? Again, it seems reasonable to suppose that she will choose the pleasant alternative. To choose pain, in full knowledge of the consequences, would be to choose to harm oneself. There can be no sufficient (self-justifying) reason for such an action; hence a choice for such an alternative must be the result of a mistake.

---

[42] See, for example, also Anscombe (1981: 30).

Hence it is being claimed that people *will* do what they *should* do, this might well seem an "over-optimistic" view of human action.

In an ethics that is egoistically grounded this type of opposition between what is done and what should be done seems to disappear. If the "good" is defined as something in opposition to the wants of the individual, then there is a motive for the person to disregard the "good". In such a case it would, indeed, be thought unlikely that all people will automatically do what is good. However, if the "good" is defined egoistically, i.e. the wants of the individual and the "good" overlap by definition, then there is no reason left for a person to disregard what is good. And, since the reward of the "good" is of an egoist nature, there is every reason for a person to act so as to attain it.

Hence the "goodness" of pleasure means that it *should* be sought, and also that it *will* be sought. Note that this is not to suppose that all choices are "consciously" geared towards achieving pleasure. Neither is it hard to point out any number of instances where people do not appear to be seeking pleasure in any sense whatsoever. This does not falsify psychological hedonism. The reason why this is so will become apparent from chapter 4 onwards. The essence of the explanation, however, is that such an action is only possible if one is ignorant of the fact that pleasure *should* be maximised, or are unable to do so, and hence rests on a factual mistake, or simple inability[43]. Hence the version of psychological hedonism in this thesis is that people always do, in some sense, seek to maximise pleasure, except if their reasoning include a mistake of some sort that blinds the actor to the fact that pleasure *should* be sought[44]. There

---

[43] The calculations and judgements needed in order to evaluate the pleasure that will accrue from pursuing certain options are most definitely not under my conscious control. "Mistakes" that result from computational processes not under my control cannot be said to count against the psychological hedonism propounded here. In Ainslie (2001) a theory of motivations in which the reward of action is discounted hyperbolically, instead of exponentially, is presented. Such hyperbolic discounting leads to inconsistent preferences, but is argued to be adaptive (2001: 45-47). If desires are seen as deriving from estimations as to potential pleasure, then this means that these estimations, "programmed" by evolution, rely on computational processes that are biased in favour of short-term pay-offs. In chapter 6 it is argued that the agent is manipulated by evolution *via* preferences. If this is correct, then Ainslie's hyperbolic discount curves form another *locus* of such manipulation.

[44] There is an important consideration that speaks against the persistence of such a mistake. If an actor (by mistake) seeks to pursue a goal that clashes with pleasure maximisation, then she will suffer the "unmistakeable punishment that is a punishment in-itself" – pain. Hence there will be an ever-present and unmistakeable incentive to abandon such a course of action. The converse hold true for actions that produce pleasure, they gain the reward that is "in-itself rewarding". This loads the dice in favour of actions that produce pleasure, and, in the long run, renders habits that do not maximise pleasure fundamentally unstable. (This is clear enough in the case of actions like touching a hot stove-plate.

are other considerations with regard to our actual cognitive capacities that makes the above statement less crude, but these considerations will have to wait till chapter 5.

Hence the position taken in this thesis will be that ethical hedonism implies, yet logically precedes, psychological hedonism.

6. Conclusion

In chapter 1 the outlines were drawn of a theory of cognitive, egoist, ethical hedonism. This theory, in a nutshell, states that individuals should maximise value (as defined), and adds that this statement should be understood as cognitive. In this chapter some issues regarding the various components of this theory were clarified. It was also argued that it stands up well against certain historically influential objections to such a theory.

The claim that people both should and do maximise value gives rise to a bewildering variety of questions. The rest of this thesis will attempt to examine some of these questions, and to show how they can profitably be dealt with.

The first question that will be discussed relates to the conceptualisation of "value" in other academic fields. No discipline stands in greater need of a conception of "value" than economics. The next chapter will examine how economics has dealt with this question, both historically and currently. The ways in which economics deals with question of value will also be related to the theory of ethical hedonism being proposed in this thesis.

Actions that cause pain less immediately and directly complicate the picture, but still leaves the difference between them and actions that cause pain directly and evidently one of degree.)

Chapter 3: Value and Economics

1. Introduction

Few academic disciplines would seem to have a more obvious need for a theory of value than does economics. Trade, production, etc. are activities that try to gain something *of value*, whether it be from man or nature. Hence it would appear that some sort of understanding of value would be a logical starting place for economics analysis.

It also makes intuitive sense to suppose that, if there is anything to the analysis of value developed in chapters 1 and 2, this must in some sense relate to "economic phenomena". This chapter will attempt to steer a dual course: on the one hand it will attempt to relate the foregoing analysis to "value", as understood in academic economics. On the other hand it will attempt to relate the "value", as conceptualised in the previous two chapters, to a sort of rough, qualitative understanding of basic economic phenomena. The problem with such a dual course is that it might well end up satisfying no-one. A philosopher might well not see the point of a detailed consideration of some economic concepts. An economist might well think an analysis in terms of value-as-pleasure somewhat quaint and outdated. Here I ask for the indulgence of the reader, the point of much of what is to follow can only really be stated after the specific issues have been discussed.

It is customary, especially when discussing the classical economists, to draw a distinction between "objective" and "subjective" theories of value. Both of these types of value-theory identify a concept that is distinct from price, and yet determines it. In an objective theory this concept relates to some quality of reality that is independent of the economic agent, in the subjective theory this concept refers to something agent-relative.

One of the key moments in the history of economics occurred when the agent-independent theories of the classical economists was replaced by the agent-relative theories of the Marginalists. The "Marginalist Revolution", as economic historians

like to refer to it, played a large part in making economics what it is today[45]. This is especially interesting, seen from the point of view of this thesis, since the particular agent-relative theory of value that was typical of the Marginalists was mostly hedonism, or something close to it. It must surely speak in favour of hedonism that its acceptance in economics allowed for insights that would probably have remained obscured for much longer otherwise.

But economics' allegiance to hedonism, though fruitful, was only temporary. The theory of value became increasingly formalised, and was correspondingly purged of psychological and hedonist elements, until only the very essence needed for economics remained. Hedonism proved to be a Wittgensteinian ladder that, once it had served its purpose, could be kicked away at no cost to economic theory. In this manner the economic theory of value, or utility, is often said to have developed from a "psychology of decision-making" to a "logic of choice".

The above did not amount to a rejection of hedonism as such. Rather than finding hedonism to be false economists simply discovered it to be unnecessary to economic theory. One thing that is necessary, though, is that the economic theory of value be agent-relative. It is this notion of agent-relativity that is probably the main historical contribution that hedonism made to economics. In this manner economics can be said to have moved from the Classical rejection of hedonism, to the Marginalist acceptance of hedonism, to the contemporary formalised notion that would be consistent with hedonism.

This chapter will start by giving a rudimentary sketch of the relation between value, as defined in chapter 1, and economics. It will then show attempt to show how vital the difference between considering value to be agent-independent and agent-relative can be for our view of economic behaviour. The arguments put forward by Smith and Marx in support of their Classical agent-independent theories will then be considered. It will be shown that these arguments are clearly inadequate.

---

[45] Schumpeter considered Walras – one of the co-inventors of "Marginalism" - to be the greatest of all economists. Of the Walrassian system, in part only possible because of the agent-relative conception of value, he writes: "It is the outstanding landmark on the road that economics travels toward the status of a rigorous or exact science, and though outmoded by now, still stands at the back of much of the best theoretical work of our time" (1994: 857).

The next issue to be considered will be the Marginalist and hedonist theory of one of the originators of the Marginalist Revolution, Jevons. It will be shown how the hedonist conception of value provides for economics to be entirely based on the concept of value. Yet the treatment of hedonism by Jevons and his contemporaries still contains some conceptual confusions. These will be shown and discussed.

The chapter will conclude with a consideration of the contemporary, formalised, conception of value as used in "revealed preference theory". It will be argued that this method can sometimes get rid of intentional description, but not always.

The first issue to be considered is the relation between the foregoing chapters and "price".

2. Value and Price

2.1.1. Definitions - Utility and Price

The following definitions are needed for a rudimentary explication of the relation between value and price.

- Value/Utility: "Value" is the sum-total of cardinal "values of experience" over time. "Valuation" is defined as the subject's *estimate* of resulting "value of experience" over time. While the actual resulting "value" is a fact about the world (as argued in chapters 1 & 2), "valuation" is an estimate as to "value", as such it can be correct and incorrect to various degrees. In accordance with economic usage the use of "valuation" will be discontinued in this and the following chapter. Instead the term "utility" will be employed to refer to "valuation".
- Price/Exchange relation: The amount of a certain commodity exchanged for an amount of another commodity. This amount is also then the *price*, as expressed in terms of another commodity.

2.1.2. Definition: The "Utility of a Commodity"

The phrase "utility of a commodity" is being dealt with separately, since it contains certain subtleties and is central to what is to follow. If value is thought to be an inherent quality of an object, then the phrase "value of *x*" presents no particular problem. But, as was argued in chapter 1, the phrase "value of *x*" ("or utility of x") can only apply in this unproblematic sense to the experience of a given subject. A commodity, however, is not an "experience", but rather a *part* or *element* of this experience.

The easiest way to explain the rationale for the way in which "utility of *x*" will be understood in this thesis is by relating it to the conclusion at which this chapter wishes to arrive. This conclusion, given certain assumptions that will be stated later, is this: If A obtained *x* in exchange for *y*, then *x* must necessarily have a higher utility for him than *y*, at the time of exchange[46].

At first glance this seems reasonable, but becomes problematic if *x* was not itself the reason for the exchange. Consider: A buys a book from B for $7. It sounds plausible to say that the book had a higher utility to B than the $7. But this is not always true if we employ the "ordinary meaning" of such a statement. Employing the ordinary meaning of the phrase "utility *of the book*" someone could say the following: B happens to be a pretty girl, and A didn't really want the book more than the $7. A just bought the book, because it afforded A a chance to speak to the girl. Here the statement that A wanted the book more than the $7, is, employing the ordinary use of these terms, simply false.

To save the above statement, one option would be to, at the cost of slight inelegance of formulation, say that A bought both "the book" and the "conversation with the girl". Here we are stretching the ordinary meaning of "buying", since B may now be an "unwitting seller", but it does save the statement needed to explicate the relation between utility and price from possible falsity. A somewhat strange consequence of

---

[46] This roughly corresponds to what a modern economist would call the "axiom of revealed preference", which will be discussed below.

such a move, however, is that the content of what is "bought" must now depend on motives of the buyer.

This difficulty, though probably not the most serious, can be avoided by employing a second option, as I will explain. The basic difficulty is that A is not really choosing elements of experience independently of one another, rather A is choosing between different experiences as such. Hence any regularity with regards to such choice must, eventually, materialise in terms of the experiences, and not their elements of the experiences.

As a heuristic I will use the vocabulary of "possible worlds". It is easier to think of these matters (for the author, at least) if an actor is viewed as choosing between the worlds that it would be possible for this actor to inhabit. This is not supposed to imply any deep points concerning modal logic, but is meant as a convenient way of writing "an option, including its consequences, and the experience to which it gives rise".

If the terminology of "possible worlds" is employed, then the principle that I am trying to formulate can be stated as follows: any chosen possible world has a higher utility, to the subject concerned, than all non-chosen possible worlds. The subject engaged in trade is now described as "choosing possible worlds" rather than "choosing commodities". Using the above example, it can now be said that A placed a higher utility on the possible world in which the book was bought and the $7 spent than on the possible world in which the book wasn't bought and the $7 kept. This applies *without reference to the specific motives of the subject*, the somewhat murky notion of what was "really bought" becomes irrelevant. Thus the statement "if A acquired x in exchange for y, A placed higher utility on x than y" should be translated as, for instance "if A bought the book for $7, A placed higher utility on the possible world that includes the bought book than the possible world that included the kept $7".

The above formulation has a further implication, which will become clear as the nature of "value", as defined in chapter 1, is considered. In accordance with this conceptualisation, saying that a commodity *has* a distinguishable value independently of a specific situation, in the same sense as saying that a tree has a certain height

independently of its surroundings, becomes a Rylean category mistake. The only "thing" that has "value" in this sense is "experience", objects don't have value independently of one another or of a mental state. The only sense in which an object can *have* "value" is inasmuch as it contributes to value of experience. This contribution will be influenced by the context (thus: other objects) in which it influences experience.

The above considerations lead to the following definition. The utility of a commodity is the *difference* between a possible world in which the commodity is obtained versus a possible world in which it is not. This difference can be expressed cardinally, as explained in chapter 1.

Consider the utility, in this sense, of, buying a car: The *value* of a car for a specific person is the *difference* between the sum-total of the value of experience that will result from buying the car, minus the sum-total of the value of experience resulting from not buying the car. The *utility* of the car is the *estimatio*n of this quantity.

Note that all events causally, but contingently related to buying the car, i.e. all events occurring in the world where the car is bought, but not in the world where the car is not bought, will now have an influence on the value of the car. The "value of a car" now becomes the "value of the car and all events causally implicated in acquiring it".

The idea of including all events causally implicated in buying a car as determining its value might strike one as odd, but a little reflection should show that this is the measure behaviourally relevant to action. Imagine A is buying a car, and knows that having a car will mean that (s)he will constantly be pestered to give lifts to friends. Now the foreseeability of being pestered to give lifts to friends will decrease the utility of the car, as defined above. In ordinary language the value of the car might be distinguished from the events causally flowing from owning it. This "ordinary meaning" can be referred to as the "decontextualised utility", this to be defined as excluding some causally related events. The measure defined above can then be referred to as "contextualised utility".

To illustrate this difference, consider the following If someone says that he hates his work, but keeps on doing it, then this "hate" can only be understood in the decontextualised sense. In the contextualised sense of value the work stands in a causal relation to his paycheck, and, in this sense, the fact that he keeps doing it means that it cannot have negative utility.

It is contextualised utility that will be behaviourally relevant to action. Consider the previous example regarding the car: the more of a problem A considers being pestered by friends to be, the less likely he will be to buy the car, or, the less he will be willing to pay for it. In this way the events causally related to buying the car impacts upon his decision to buy the car. This should demonstrate the virtue of including all events causally related to acquiring a commodity as influencing the utility of the commodity itself.

A formulation that includes causal events also allows for the phrase "utility of money" to be used without it being troublesome. If A buys a book for $7, as in the previous example, that implies that the book had higher utility for him than the $7. Talking about the "utility of $7 " might sound odd, since money does not normally have utility in the same sense as other commodities. But, in the above example, it might as well have been said that the book had higher utility bought than unbought. Whether the possible world in which the book wasn't bought is referred to as the world in which the $7 was kept, or the book unbought or A had a conversation with the girl, or whatever, is immaterial. The fact that someone has $7 in a possible world is very likely to impact on the value of the world, not distinguishing between commodities and events causally related to owning commodities obliterates any relevant distinction between money and other commodities. The amount of money/currency is now *just another element in a causal chain* by which this causal chain can be named.

3.1 Simplifying Assumption

In order to simplify the discussion to follow, the following assumption will be used. Any subject weighs the possible consequences of action in terms of their value, as defined. This measure then determines action.

The above might appear quite unnecessary, or a simple restatement of chapter 1. Here a distinction should be made between the *ultimate goal of action*, and *the strategy* that best implements it. In chapter 4 it will be argued that, while all people should maximise value, this is, in itself, a poor guide to action. In other words, any attempt to maximise value by consciously using the rule "maximise value" will be significantly less effective than certain other strategies, to be discussed in chapter 4. It will also be argued that this problem leads to quite a complicated relation between "maximise value" and the strategy that implements it. It will be argued that this relation is complex enough that there is a *qualitative break* between a value-maximiser and a subject who follows a strategy flowing from it[47]. This break is of such a type that there is one important sense in which it is misleading or even false to say that people maximise value, or are psychological hedonists.

The above consideration, while it will prove to be vital to the rest of this thesis, will complicate the analysis and discussion in this chapter unnecessarily. Hence, for this chapter, it will be assumed that people are value-maximisers in the simplest sense.

There is a further, less troublesome assumption that is needed. In the examples under discussion the value of a commodity will be treated as a function of the amount of that commodity. Hence it is being assumed that, all else being equal, more of x has higher utility than less of x[48]. It need not be assumed that this increase is uniform.

3.2 The Relation between Utility and Price

The above considerations and definitions allow for a clear explication of the relation between value and price. Consider the case of A trading a determinate amount of a given indivisible commodity x with B, for another commodity y. What will determine the price (amount of y needed to effect trade) of this commodity?

A will trade at all prices where the trade is to his benefit. Hence A will seek to be in a possible world of higher utility than if no trade had occurred, if this situation does not

[47] This forms the crux of the debate between "rational choice"-theory and "bounded rationality", it is in these terms which the issue will be discussed in chapter 4.
[48] Economists call this the "principle of non-satiation".

obtain A will not trade. The most advantageous trade, if the commodity which is being asked for as a price is the only variable, is one which occurs at a price of zero. This would, in other words, be where A manages to obtain commodity *x* without surrendering any of the commodity *y*. All trades at a price higher than this are *progressively* of lower utility to A. If A would buy at a price of zero, progressively increasing the amount of the commodity *y* (hence: price) would lead to a situation where eventually A is indifferent between trading and not trading. Here the possible worlds of trade/don't trade are of equal utility. This price in a world where A is indifferent between trading and not trading is the maximum price, at all prices fractionally less A would trade. In this way utility, as defined above, determines the maximum price at which a commodity will be traded. If a trader is treated as being both a buyer and a seller, then the converse is true of the trader *qua* seller. Utility now determines the minimum price at which the seller would be willing to trade. Hence the utility of the given commodity x sets the maximum price for the buyer, and minimum price for the seller. This implies that utility determines a *range* of possible prices[49].

Hence the relation between utility and price can be stated very simply: the utility of a commodity for the buyer at a price of zero determines the maximum price for the buyer, and the utility for the seller determines the minimum price for the seller. This utility is the *difference* in estimated value between the possible world in which trade occurred at a price of zero, versus a possible world in which trade did not occur. This utility determines the range of possible prices in the sense that a determinate amount of the commodity offered as price would be needed to "close the gap" between the world in which price is zero and the world in which buy/don't buy are of equal utility.

3. Agent-Relativity vs. Agent-Independence

3.1 Trade is not Zero-sum

The above analysis renders the first objective of this chapter achieved. Value has now, *via* the idea of estimated value ("utility"), been related to price. The above analysis,

[49] Under extra assumption price becomes determinate, as will be discussed below.

however, being a rather straightforward application of hedonist philosophy, might well seem trivial This is, however, not the case It is important because it establishes that trade is not a zero-sum activity. It is also of historical interest, since some major figures in economic history (Smith, Marx) thought it to be false. These two issues will be discussed in turn.

If value is treated as an agent-relative notion (as it is in hedonism), then both parties to a trade can gain. If I can increase utility by buying a cup of coffee, and a restaurateur can increase his utility by selling it to me, we both win. Indeed, such a structure of mutual gain is the only possible explanation of any voluntary trade whatsoever. A trade that is not in favour of both is only possible if one party had, somehow, miscalculated.

The above is a clear consequence of having an agent-relative conception of value. An agent-independent view of value changes the picture in a disturbing way. Now a commodity must, in some sense, have a real value. This one real scale of value turns trade into a zero-sum game. Here trade now becomes a game in which one party must win, and another lose by the amount that the first gained. The only other possibility is that the good must somehow trade at its real value, and neither gain or lose on this objective scale[50].

The contrast between the above two views cannot be greater. If someone has an agent-relative view of value, then trade is a desirable activity that makes society better off, and should be encouraged. If a commodity has an inherent value, however, it becomes significantly less clear that trade is a virtue. Take the example of seeing a rich person walking down the street. If I have an agent-relative conception of value, and assuming that the person didn't steal or inherit her money, then it is *prima facie* more likely that I will consider the person socially useful. If someone gets rich by trading with people who didn't systematically miscalculate utility, then the wealth of the person becomes a reflection of their use to society. Simply put, the person might well be rich because an awful lot of people have found it in their own interest to have some form of economic interaction with them.

[50] Marx's distinction between the use-value and exchange-value of labour allows him to avoid this otherwise inevitable implication of having an objective theory of value.

The situation is much more sinister if I have an agent-independent conception of value. Now the riches of the person must exactly correspond to the amount that others have lost in trading with them. In this sense there is now something inherently *exploitative* in the very idea of gaining from trade. Hence it would be more natural to see the person as a drain on society, and to see wealth as a badge of shame.

What does the above show? It is not intended as a simple-minded apology for capitalism or the rich, etc., there are a host of issues involved in such arguments that have not been discussed here. Nor does it claim that an economist's view of value is all that will determine her views on capitalism[51]. What it does try and demonstrate is how much of a *prima facie* difference the conception of value as "subjective" or "objective" can make for our intuitive understanding of economic phenomena. The idea of mutual gain is the *determining feature* of the one view, whereas it is a *logical impossibility* in the other.

The agent-relative view of value, as developed above, permits an explanation of trade without ever postulating the idea of the "value of a commodity", in the literal sense. All there is, fundamentally, to economic phenomena is the agent-relative conception of value and the different prices to which this can give rise. The phrase "value of a commodity" is simply meaningless if it is supposed to refer to anything other than this.

Classical economists found reason to doubt the agent-relative view of value. It is the arguments supporting this doubt that will next be discussed.

3.3 Adam Smith's "Cost of Production" Theory of Value

Adam Smith's doctrine of the "invisible hand" (Smith, 1975: 400) relies on the mutual beneficence of trade and thus on an implicit agent-relative conceptualization of value. And yet he dismissed utility *as a determinant of price*, he separated the two

[51] Smith, despite having an agent-independent theory, was pro-capitalist. Walras, despite an agent-relative theory, was a socialist. (Smith does however, depend on an agent-relative view to explain his idea of the "invisible hand".)

in order to advance a cost of production theory of value with labour as an independent[52] determinant of price. How is this accomplished?

> The word value, it is to be observed, has two different meanings, and sometimes expresses the utility of some particular object, and sometimes the power of purchasing other goods which the possession of that object conveys. The one may be called "value in use", the other, "value in exchange". The things which have the greatest value in use have frequently little or no value in exchange: and, on the contrary, those which have the greatest value in exchange have frequently little or no value in use. Nothing is more useful than water: but it will purchase scarce anything; scarce anything can be had in exchange for it. A diamond, on the contrary, has scarce any value in use; but a very great quantity of other goods may frequently be had in exchange for it (Smith, 1975: 25).

This is the classical formulation of the diamonds-and-water paradox[53], a problem that strikes at the heart of Classical economic theory since it can only be resolved with reference to a theory regarding the determinants of price. Adam Smith (and other classical economists like Ricardo) uses it to divorce utility from price and clear the way for an "objective" theory of value that assigns prime importance to labour. The need for such an "objective" theory does not arise if this paradox can be solved and utility related to price.

Utility has already been related to price earlier in this chapter, but how does this relate to the diamonds-and-water paradox? Like most paradoxes, it should be shown not to arise, rather than be "solved".

Smith denies the claim that the high price of diamonds can be accounted for by their high utility, he states that that they have scarce any "value in use". This is a puzzling statement, since diamonds are very useful things. They are used, both industrially and decoratively, the latter in virtue of the great symbolic power they have in our society (and Smith's). His declaration that they have "scarce any value" is probably best

---

[52] In a primitive society without capital and the private ownership of land, he makes labour (quantified as labour-time) the sole determinant of the exchange relation (price) (Smith, 1975: 42).

[53] O'Brien notes that Smith "solved" the diamonds-and-water paradox in his *Lectures* in terms of relative scarcity, but changed his position in *The Wealth of Nations* in order to advance a "cost of production" theory of value (O' Brien, 1975: 78-80).

interpreted as the result of a bias against their decorative use. Whether society *should* value something is irrelevant to an explanation of the prices of objects, it is not the task of the economist to judge the tastes of mankind.

To indulge in a bit of speculation, I think it is plausible to assert that the idea of objects having a "real" agent-independent value might be to blame for this. One might well be tempted, in thinking in such a manner, to draw a distinction between uses that are only possible in virtue of the nature of the object, and the uses that "are added" on by social custom and convention. Such a distinction would then lead to the temptation to say that diamonds are not "really" valuable. Whether such a distinction can be coherently maintained will not be discussed here, the important point is that it is irrelevant to an explanation of prices in a given society. In our society diamonds are, among other things, markers of class, status and affection. As long as they have these functions they will be very useful things to have.

Smith's statement that diamonds have little value is, in this context, simply incorrect. Whether utility is, in part, the result of social convention, does not matter one iota in the determination of prices. Since diamonds have great utility, and since utility sets the upper and lower limit for prices, diamonds constitute an instance of the claim that price determines utility, not an exception to it. This is, in a sense, not the whole story, for the fact that diamonds fetch high prices is not independent of the fact that they are hard to obtain. This is the phenomenon that allowed Smith to assert that ultimately, cost of production trumps utility in the determination of prices. The difficulty of obtaining a commodity will be related to prices later on when the Marginalists are discussed. For now it will have to suffice to have shown that the "diamonds" part of the "diamonds-and-water" paradox does not present a problem.

The "water"-part is similarly unproblematic, water has a low utility and hence a low price. The previous statement might sound paradoxical, for wouldn't we perish without water? This issue is hard to discuss without, to a certain degree, mentioning issues regarding the relation between utility and the difficulty of obtaining a commodity. This topic will be discussed later, it will only now be mentioned to the degree necessary to make the above statement seem less strange. This will be easiest to do with reference to a hypothetical case which illustrates the principle under consideration.

Consider two people on an island which contains a river with fresh water. Imagine A regularly goes to the river for water, and that B regularly trades with A in order to obtain some of her water. What determines the value of the water for B, and hence the amount of the items he will be willing to trade?

If utility is seen as the *difference* between two possible states, then the possible state that does not materialize has a definitive influence on utility. For B the value of the water is the difference between the possible world in which it is obtained, versus the possible world in which it is not. Herein lies the solution to problem. If B does not obtain the water from A, then surely nothing prohibits him from getting it from the river himself. Hence the value of the water obtained from A is the difference between the possible world in which it is obtained from A and the possible world in which it is obtained from the river.

These two possible worlds do not differ with regards to the actual commodity, i.e. the water. Since B acquires water in both worlds, the difference between the two possible worlds, and hence utility, cannot be ascribed to any of the uses of water as such. Rather it must be ascribed to some other elements in the causal chains which is being named by "utility of the water". Clearly one rather sizable difference is the difference between obtaining the water from A and going to the river. The greater the disparity between these two, the greater the utility of the water. The utility of buying water from A is the difference between a possible world in which it is obtained at no price ("price" here meant as the specific variable used to effect trade), and a possible world in which B has to go to the river himself. Provided the river is close by, B is healthy, etc, this difference will not amount to much. Hence A's water will have a low utility, which results in it having a low price.

The key thing is again to be suspicious of Smith's of phrase "utility of water". For matters of trade there is no such thing as the utility of water as such, there is only the utility of this-water or that-water. And, since utility is a difference between possible states, the this-water of a possible world is a commodity in competition with the that-water of another possible world, and hence can influence its utility.

It is true that, without water as such, we could die of thirst. But water as such is never the object of trade, all that is sold is some specific amount of it. If someone could monopolise the water-supply of the world, and hence remove the possibility of obtaining water from the lower of the two possible worlds, then Smith would be correct. Here the water of the person would have a great value. But the paradox would again be averted, since it would also then have a great price.

The above way of showing that the diamonds-and-water paradox does not arise raises the question of the relation between the conditions of obtaining a commodity and utility. Someone versed in the history of economics will also notice that the above is not quite the traditional solution to the diamonds-and-water paradox in terms of the "law of diminishing marginal utility". These issues will be easier to discuss below in the context of the Marginalists' work.

Hence the diamonds-and-water paradox rests on conceptual confusion. If stated in terms of the measures of utility that determine price, diamonds have a high utility and water has a low utility. Smith's reason for separating utility and price, and advancing a cost of production theory of value, fails.

I would contend that the conceptual confusion behind the formulation of the diamond-and-water paradox is probably due to being mislead by the phrase "value-of-a commodity". Although such an assertion is necessarily speculative, it does seem that Smith's intuitive conception of use-value is something that is independent of society (diamonds) and conditions under which objects are obtained (water). Such a view seems best accounted for by ascribing to Smith the belief that the usefulness of objects is something inherent in these objects, and not a relation between the qualities of the object and the needs served by these qualities[54].

Another famous "objective" theory of value is the Marxian Labour theory of value. Marx does not use the diamonds-and-water paradox to separate utility from price, but

[54] In a pre-capitalist society, Smith makes labour the sole determinant of price (1975: 42). His argument for separating utility from price was criticised above. He does not seem to really present an argument for the next step of identifying labour with price. Smith only writes that labour controls price because it is "natural" (page 41), that labour is the "real price" (page 26), that labour is the "real cost"

still needs an argument to accomplish this in order to clear the way for labour as determinant of prices. It will be shown how his argument works and why it fails.

3.4 Marx's Labour theory of Value – The Argument from Incommensurability

Marx's economics rests largely on his theory of value[55], which is generally conceded to be incorrect[56]. Yet the beginning of his argument is of interest in terms of the ideas developed in this chapter.

Marx begins by distinguishing between utility and price in a way reminiscent of Smith. He then proceeds to separate the two by saying the following:

> As use-values, commodities differ above all in quality, while as exchange-values they can only differ in quantity, and therefore do not contain an atom of use-value (Marx, 1976: 128).

When Marx says that use-values differ above all in quality, he means that there are different kinds[57] of utility, and while they can be quantitatively compared for a given use-value (e.g. different amounts of shoes, buildings, etc.), they cannot be quantitatively compared for different use-values. Obviously exchange-value (price) is quantitative, and this incommensurability between utilities must mean that utility cannot be the commensurate entity controlling prices. From here Marx searches for this commensurate quantity, and eventually announces it to be the amount of labour-time socially necessary for the production of a commodity. Again, as with Smith, if we can show that the above argument from incommensurability is unsuccessful in separating utility and price, that means that the need for Marx's "master-concept"

(page 26), the "first price" (page 26), the "original purchase-money" (page 26) and something's "real worth" (page 26).

[55] In a letter to Engels he writes; "...(t)he best points in my book are: (1) the twofold character of labour, according to whether it is expressed in use-value or exchange-value. (All understanding of the facts depends upon this.) It is emphasized immediately, in the first chapter..." (Mclelland, 1977: 525).

[56] The *New Palgrave: Marxian Economics*, asserts the Labour theory of value to be false on the opening page without feeling the need for argument or even the need for a citation to back up this claim (1987: xi). (In a similar way, even the greatest admirer of Adam Smith will not locate his considerable merits in his theory of value.)

[57] Consider page 136, where he writes that "...[with regards to use-values] it was a matter of the "how" and "what" of labour, [with regards to exchange-values] of the "how much"...". Or, page 126: "...use-values of one kind exchange for use-values of another kind".

(labour-time) does not arise and that his analysis is unsupported by the time labour gets mentioned in the first pages of *Capital*.

It should be reasonably clear that the argument from incommensurability is irrelevant to the conception of utility developed earlier in chapter 1. All objects can have an influence on the value of the experience of the subject. This influence is a determinate cardinal[58] quantity and thus commensurate. But this does not correspond to Marx's definition of utility, he writes that:

> The usefulness of a thing makes it a use-value. But this usefulness does not dangle in mid-air. It is conditioned by the physical properties of the commodity, and has no existence apart from the latter. It is therefore the physical body of the commodity itself, for instance iron, corn, a diamond, which is the use-value or the useful thing. This property of the commodity is independent of the labour necessary to appropriate its useful qualities (Marx, 1976: 126)[59].

It seems clear that Marx views the use-value of a commodity as an inherent quality of it based purely on objective factors. On his definition, use-value cannot be quantitative. If we fall into the trap of thinking that the above definition exhausts the possible conceptions of value, then we are committed to admitting that utility cannot influence price.

But there are other conceptions of "utility" or "use-value" which do not fall prey to the problem of incommensurability. One such conception is the hedonist one developed in this thesis. Marx does not consider such conceptions, nor does he present any argument against them.

---

[58] The claim of cardinality would not be needed to defeat the idea of incommensurability. Any conception that can give rise to an ordinal ranking of preferences will do.

[59] This passage is not easy to interpret. Marx says that use-value is both "conditioned by the physical properties of the commodity" and is "the physical body of the commodity itself". This appears to be a flat contradiction. Nothing can be conditioned by something else, and be identical to that something else, at the same time. Regardless, it seem clear that Marx is trying for a conception of "use-value" that is as "materialistic" and "objective" as possible.

Marx's argument trades on defining use-value in such a way that it cannot be quantitative, implicitly assuming this to be the only possible conception, and then using this definition to dismiss use-value. This is done on the grounds of not being quantitative (and thus commensurable). This argument says more about the peculiarity of his definition than about the influence of utility on price.

Although it has already been argued that Marx's argument from incommensurability is fallacious, and hence his argument for his theory of value unsupported, it is instructive to see how this argument develops further. On the opening pages of *Capital* he writes that, if two commodities are traded, this implies that "a common element of identical magnitude exists in two different things"(1976: 127). This "common" (commensurate) element will be found to be labour-time (128).

The idea of a "common element" has been discussed above, the idea that this exists in "identical magnitude" should give pause. Marx does not argue for the above idea, he asserts it to be self-evident that trade is only possible between commodities that are, in some sense "equal". In the rather rudimentary explication of the relation between utility and price given earlier in this chapter it was shown that the idea of equality between traded objects is nowhere needed to explain a given trade. Indeed, if there is anything to the idea of an agent-relative conception of value - and it has served economists well for over a hundred years - then the idea of equality between traded objects is not only not self-evident, but a conceptual impossibility. There seems to be no other way to explain Marx's statement, and the careless way in which it is made, as the result of uncritically accepting the idea that the phrase "value of a commodity" has a quite literal sense.

With these challenges to an agent-relative conception of value having been discussed, it is now time to look at the theory of value that supplanted that of the Classical economists. This is the theory of Marginal utility.

4. The Marginalists

4.1. The Basic Marginalist Idea and its Relevance

The Theory of Marginal Utility was originally proposed by Gossen in 1854. It was mostly ignored until independently formulated by Jevons (1862 in outline, 1871 in full), Menger (1871) and Walras (1874)[60]. Once accepted, it signalled the end of Classical economics and the rise of the Neo-Classical school; the main difference lies in the rejection of the labour theory of value and it's replacement with a "subjective" theory of value based on an agent-relative, and hedonist, theory of utility.

The basic idea behind the theory of marginal utility is not dissimilar to the analysis carried out earlier in this chapter. It does contain certain extra assumptions and idealisations in order to arrive at more elegant and manipulable answers. The following are the most important.

- The law of diminishing utility: For every added unit of a commodity, the total utility ("pleasure") of the commodity increases, while the additional utility decreases uniformly. In layman's terms, that means that I like the second Coca-Cola less than the first, the third less than the second, etc.
- That trade ceases before the stock of buyer or seller is exhausted.
- The infinite divisibility of commodities.

The reasoning underlying the doctrine of marginal utility is very simple, and rests on the fact that no person will trade at a disadvantage to himself. If A and B are trading salt for pepper, the trade can start at some arbitrary exchange relation (price), for instance one handful for two handfuls. If A is the person originally in possession of salt, every unit of salt that he gives up is progressively more valuable, and every unit of pepper that he gains is less valuable. Once the last unit of salt is equal in utility to the last traded unit of pepper he will cease trading, since any further trade will be at his disadvantage. Whereupon B, if he still wishes to trade, can change the exchange

[60] See Kauder (1965) for an excellent account of the history of the theory. Another excellent introduction can be found in Ross (1999). Ross also discusses the period after Marginalism in an accessible manner.

relation by offering more pepper per unit of salt. It is now again to A's advantage to trade, and trade will continue until one party again finds that his last traded unit are equal in utility. Now the other can again entice him to trade by changing the exchange relation. This process will cease once both parties find that their last traded units are equal in utility and thus stop at the same time. They are now said to be in equilibrium.

This means that the marginal utilities of both parties individually are the reciprocal of the exchange relation between the two commodities, if this weren't the case then trade would continue. This also means that a consumer buying a given product in increments will keep buying until the last commodity bought has the same utility as the money surrendered. Thus, at equilibrium, marginal utility is equal to price.

The above reasoning is consistent with the analysis made earlier in the chapter[61]. But, instead of utility setting a range of possible prices, price is now exactly equal to marginal utility due to the extra assumptions. This allows for the elegant mathematical description of the conditions under which economic systems will result in equilibria, as well as for understanding as to the way certain factors in such an equilibrium interrelate in order to achieve these equilibria[62]. This work was most comprehensively carried out by Walras - which is why Schumpeter singles him out as the greatest of economists.

The Marginalists explicitly conceived of utility as a real entity, to be equated with what we call pleasure and pain[63]. This allowed for the construction of models and the derivation of theorems which still, in some form or other, lie behind much of modern economics. This, however, is not to say that modern economics is necessarily hedonist. Economists later discovered that these ideas, discovered by way of the hedonist approach, can also be arrived at without making any such assumptions as to human motivation. This was done by stripping the concept of utility from psychological and hedonist associations down to the bare minimum needed for

[61] Except for differences in terminology, the analysis made earlier in this chapter is also made by Jevons (1911:118-127) himself. Here it is done for the special case of indivisible commodities.

[62] It also marked a leap forward in the mathematization of economics. Jevons' preface to his *Theory of Political Economy* is also a historically fascinating manifesto for the increased use of mathematics in economics.

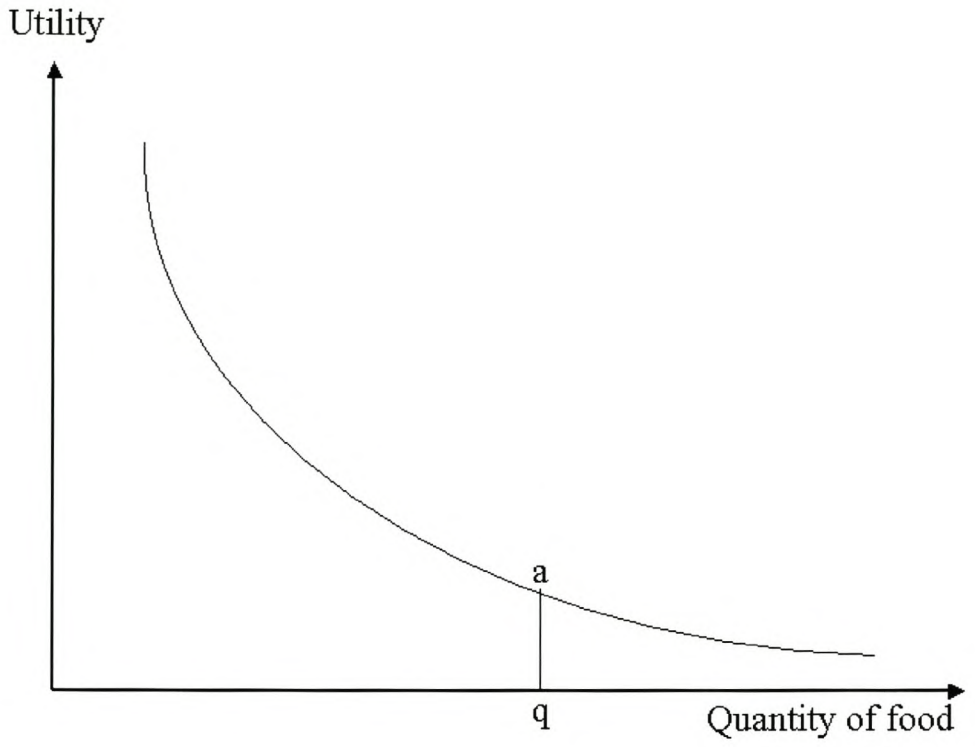
economic theory. This achievement ranks as one of the greatest in economic history, and will be discussed below.

First, however, it will be instructive to see to what degree the Marginalists managed to purge themselves of the conceptual confusions regarding value that seem to lie behind some of the statements of Smith and Marx. This will also present the opportunity of extending the analysis made earlier in this chapter by discussing the determinants of utility. This will be done mainly with reference to Jevons, for no reason greater than the fact that he was, in addition to being an economist, also a philosopher. The differences between his version and the other contemporary versions of the theory arise at a level well beyond what is necessary for the discussion of this chapter.

4.2 The Marginalist Conception of Utility

Jevons, after explicitly stating that his work is built on Benthamite foundations (1911:28), considers the food eaten by a person in a day, and analyses the utility of the successive portions. The essence of his analysis is contained in the graph below, where the y-axis represents utility and the x-axis the amount of the commodity:

[63] They were, however, not all hedonists in the full sense. Some believed hedonist motivation to be true only in "lower" matters like economic activity, but insufficient to account for all behaviour. See Kauder (1911: 93-97).



The following should be noted:

- The y-axis is not touched because the first increment has "infinite utility". This implies that total utility is also infinite This "infinite utility" is due to the fact that food protects from starvation (Jevons, 1911: 45).
- The utility per unit does not decrease relative to a possible world, but just decreases "in-itself".
- Price is determined by marginal utility, which is quite low and represented by line *aq*.
- Utility is a fundamental datum, and decreases as a function of the amount consumed. This is due to the law of decreasing marginal utility.

The above "marginal utility curve" contains the essence of the revolutionary ideas behind marginalism[64]. Yet it still contains the result of the type of conceptual confusion found in Smith and Marx.

Jevons' does not allow the curve to touch the y-axis, since the total utility of food is infinite. This seems similar to Smith's idea that water has great utility because it is needed to maintain life. It was discussed above why this is incorrect, the fact that water as such maintains life does not confer any great utility on any particular concrete instance of it. In a similar way Jevons' idea that the first increment of food consumed (and hence the total food consumed) protects from starvation is, as far as the measure relevant to price is concerned, fallacious[65]. For, if the consumer cannot purchase any particular "first exemplar" of "food", then surely nothing prohibits him from buying another "first exemplar" of "food". And, since utility is a difference between states, the nature of the second state, and hence the possibility of other food, plays a large role in determining the utility of any particular piece of food.

Jevons' definition of utility is a standard hedonist one (1911: 38-39), yet he does not explicitly state that the utility of any particular commodity should be seen as a *difference* between possible states. Yet the assumption that it is a "difference", and hence that what happens if we do not get something helps to determine the value of that something, is needed if he wishes to say that food is valuable because, without it, we starve. But the full implication of this is not realised by this very same statement. For if utility is a difference, and an attempt is made to determine the utility of any specific quantity of food, then *all other food in the world becomes a potential replacing commodity*. Again, as with Smith, he seems to be under the spell of the idea of the value of food as such, yet food *as such* is never traded, only parts of it.

Jevons and his fellow Marginalists chastised the Classical economists for a failure to distinguish between the total utility of a commodity and the utility of any particular part of it. Such failure was thought to be the source of paradoxes like the one

---

[64] Something roughly like it is still used in textbooks today for heuristic purposes, as it provides an intuitively plausible account of the relation between the amount and utility of a commodity.

[65] The false statement that the first increment of food protects against starvation can be made analytically true by taking "first increment" to refer to whatever piece of food I eat first. In such a case, if I do not have a first increment, then eating a second increment is impossible by definition. Here a distinction needs to be drawn between "a piece of food which happens to be the first increment and hence can be replaced by another piece of food which will then happen to be the first increment" and the "first increment" proper. It is possible that Jevons does not quite distinguish between the two at all times. I am interpreting Jevons as using the first definition throughout, since if he used the second throughout his analysis would have no bearing on economics whatsoever. "First increments", in the sense that would make Jevons statement trivially true, are not what are bought and sold on a market, rather a large amount of potential "first increments" are bought and sold.

concerning diamonds-and-water. It was already shown, above, that the distinction between total and marginal utility is not necessary to dismiss the diamonds-and-water paradox. Rather the problem lies with not fully realising the implication of seeing utility to be a difference between possible states, and not a quality of an object in a state. This is ironic, since the same conceptual confusion that led the Classical economists to the problem involves Jevons and company in the same paradox, except that they can solve it by relating price to marginal utility[66], instead of utility.

In the great founding work of modern economics, Smith's *Wealth of Nations*, he frequently interrupts his analysis for "digressions". These digressions sometimes run over a hundred pages. There are even occasionally digressions within digressions. The above remarks regarding Jevons and his fellow Marginalists pose interesting conceptual questions regarding the "diamonds-and-water" paradox. Mainly because the topic has been raised by the remarks made above, but also because they are somewhat interesting in themselves, I ask the reader to allow me one such digression.

4.3 A Digression: The Diamonds-and-water Paradox

The diamond-and-water paradox was a central problem for the classical economists. Smith (and Ricardo) use it to dismiss utility as a determinant of price and establish an objective theory of value. Yet any number of authors prior to Smith had resolved the problem by making price depend on utility and scarcity (Schumpeter, 1994: 301). Smith himself occasionally availed himself of this solution (see footnote 8) , yet he did not use it in the *Wealth of Nations*.

The problem with stating price to be determined by utility and scarcity is that this is not a full solution. The author who uses it must still be able to explain why such disparate elements are relevant, and more importantly, how they interrelate. In other words a full solution needs to explain why, when there is less of a thing, it will tend to cost more, and not just assert that it does. This was accomplished by the Marginalist "law of diminishing utility". The idea that more of a thing will result in a lower price was seen as the result of the common psychological observation that any additional unit of a good will be used to satisfy a less pressing need than its predecessor. In this

[66] Modern textbooks that mention the paradox tend to give the Marginalists the credit for the first definitive solution (Samuelson, 1989: 455).

way, if there is a lot of something, we can be sure that the last few available units aren't of great use relative to the first. And, since price is equal to marginal utility, it is precisely the last few units that determine price.

The first problem with the above is that, while marginal utility is equal to price at equilibrium, it does not *determine* it. There is no way of determining the marginal utility of a good prior to and independently of a given exchange. Rather marginal utility is itself a function of the price at which a good can be had. In this way marginal utility and price co-determine each other. And yet this does not seem to bar the "law of diminishing utility" from being an explanation of "why more costs less". It seems reasonable to suppose that if two states are identical, except for the fact that all people in one state have more of a given good than the same people do in state two, then the good can be expected to trade for less in the first state. Such a lower price can then coherently be ascribed to the law of diminishing utility.

I would contend, however, that while the law of diminishing utility can sometimes explain "why more costs less", it does not solve the diamonds-and-water paradox. This is because there is no paradox to solve, conceptual clarity can get rid of the problem before the need for any economic theory arises. Why this is the case was touched upon when Adam Smith was discussed earlier in this chapter. This explanation will now be developed somewhat.

The "law of diminishing utility" is not the only explanation of "why more costs less". Even if total utility increased as a function of amount in a rather erratic way, and not as proposed by the law of diminishing utility, it would still be possible to show that "more costs less".

The following passage from Bohm-Bawerk, a second-generation Marginalist and follower of Menger, contains the basic idea behind such an explanation. Bohm-Bawerk follows the general marginalist line by stating that price is determined by marginal utility, but then pauses to consider certain seeming exceptions:

> A coronation festival is being celebrated. Admission to this interesting event is by card, and the cards may be obtained gratis but they are non-

> transferrable, and are issued only on application in person. I posses one of these cards. If I lost it, I should not have to forego the pleasure of attending the festivity, but should have to put in another personal application for a card. What possession of the card therefore really means for me in that case is that I am relieved of the unpleasant burden of going to the proper office and making my request for a card (1959: 178).

In terms of the analysis carried out earlier, it should be fairly clear what is at stake in the above case. Bohm-Bawerk considers the utility of the card, and realises that, should he lose the card, he will simply replace it with another. This implies that the two possible worlds, one in which he uses the old card and one in which he obtains a new one, do not differ in any way with regard to matters intrinsic to the commodity (going to the festival). Hence the utility of the card must vest in something else, here the utility of the card is the result of the fact that possessing it means I do not have to go through the process of getting another.

The problem with the above is that making utility depend on the degree of difficulty of getting something seems to sound dangerously close to a labour theory of value. Bohm-Bawerk opposes this interpretation emphatically (1959: 179). He states that the value of a good always vests in its subjective utility, here this subjective utility just happens to consist in the avoidance of "some painfulness or troublesomeness" (1959: 179). He also states that the above state of affairs which give rise to the seeming exception will not occur very often, and that when it does it will "involve only trifling or insignificant matters" (1959: 178).

It is here where the analysis carried out earlier in this chapter allows for an interesting development of the above logic. Rather than treating Bohm-Bawerk's special case as a strange exception that can be ignored, it can be shown that something similar to it is, in fact, the general case for all commodities. Hence the logic that Bohm-Bawerk identified above is far from trivial, rather it can be extremely useful in getting a rough idea of how prices are determined.

4.3.1 Performance Disutility and Utility - Statement

Some definitions are necessary before Bohm-Bawerk's example can be fruitfully discussed. The term "performance" will be used to denote all actions necessary to acquire a given commodity. The actions needed to acquire a given commodity are normally actions we would choose not to perform if they didn't stand in a causal relation to acquiring a given commodity. Or in other words, performance tends to have negative utility or disutility, in the decontextualised sense. "Performance disutility" then means the negative utility of a given performance. This is understood as the difference in utility between a possible world in which a good is obtained without doing anything (where it "falls into my lap" so to speak) and a world in which it is obtained by whatever means.

What Bohm-Bawerk sees as an "exceptional case" can be, in a sense, generalised by making the particular aspects of his example less restrictive. The first peculiar aspect of his example is that the card to the coronation is obtained for free (or "gratis".) "For free" here does not constitute "without a cost", in the economic sense. No action can be cost-free if cost is defined as "all things given up in making a choice". To choose any possible world is to forego all others, these other possible worlds amount to a "cost', in this sense[67]. "For free" here rather means that the type of thing that is usually foregone in acquiring a commodity, is not foregone, presumably this refers to "money" as a *numeraire* for other marketable commodities. The same effect as is achieved by Bohm-Bawerk's "gratis" can be obtained by simply speaking of a possible world in which the commodity was obtained at no price. Here "price" again simply refers to the type of thing that one normally surrenders in acquiring an object, and usually constitutes one side of an exchange-relation.

The second peculiar aspect of Bohm-Bawerk's example is that an identical replacement commodity will be obtained if the original is lost. Again, this does not have to be a peculiar occurrence. For a wide variety of commodities it is possible, if I

[67] Or, as an economist would say, all choices have an "opportunity cost".

do not choose the original commodity, to obtain an identical[68] replacement. Hence I can simply speak of a possible world in which an identical replacement commodity is obtained. It will not, however, always be the case that obtaining the identical replacement is my next best option. (Indeed, this is never the case for infinitely divisible goods.) In such cases performance disutility does not equal utility as in Bohm-Bawerk's example. But, and this is the central point, there is still a relation between performance disutility and utility that allows for some insight into economic phenomena.

This relation works as follows. Consider the possible world in which the commodity at price zero is not bought, i.e. the "second best option". This possible world can contain an identical replacement commodity, or not. The question as to whether it will include an identical replacement commodity or not will depend on whether the world of the replacement commodity has higher utility than the world in which the identical replacement commodity is not bought. For the sake of clarity, let's use the following definitions:

- World 1: The possible world in which commodity x is bought at a price of zero.
- World 2: The possible world in which an identical replacement commodity (x') is obtained by whatever means (making, finding, buying).
- World 3: The possible world in which neither x nor any other identical replacement commodity is obtained.

Remember that it has been established that the utility of x is the difference between world 1 and the next best world, and that this utility sets the range of possible prices.

If world 2 is the next best world, then the utility of x is equal to, and determined by, the difference in performance disutility between x at price zero and x'. Thus this difference in performance disutility determines the maximum price.

[68] Identical to mean a good of the same *genus*, where I am indifferent between the original and the replacement, all else being equal. For example, most people would be indifferent between any two particular beers of the same brand, any two cans of corn, etc.

If world 3 is the next best possible world, then it follows that the utility of x at price zero is smaller than the difference between the performance disutilities of x at price zero and x'. Thus this difference in performance disutility represents a *maximum* to the utility of x at price zero.

It follows that, no matter which world is the relevant one to the determination of utility, utility can never rise above the difference between the performance disutilities of x at price zero and x'. This difference in performance disutilities represents a limit to utility, it must always be equal to, or smaller than, this quantity.

To simplify the rule we wish to arrive at, matters first need to be complicated further. It would be preferable to speak of the performance disutility of the identical replacement commodity itself, instead of a difference between performance disutilities. To arrive at this formulation, consider a possible world in which the performance disutility of x itself at price zero becomes progressively less. It should be clear that all this accomplishes is to increase the *difference* in performance disutility between x at price zero and x'. Hence, even if the performance disutility of x at price zero is disregarded, and we just refer to the performance disutility of x', the performance disutility of x' will still represent a limit to the difference in performance disutilities between x at price zero and x'.

Hence the following situation arises: the performance disutility of an identical replacement commodity is a limit to the difference between the performance disutilities of the first commodity and the identical replacement commodity, which is an upper limit to utility. Or, using the simplification explained above, *no commodity can ever have a higher utility than the performance disutility of its identical replacement.*

4.3.2 Performance Disutility and Utility – Relevance

The above reasoning is useful, in that it enables one to gain some insight into the determination of utility. There is a sense in which utility is a fundamental *datum*, and questions regarding more or less utility are the province of psychology rather than economics. But the above formulation circumvents this problem. If the utility of a

good can never be higher than the performance disutility of its identical replacement, then an understanding of changes in performance disutility can help towards an understanding of changes in the magnitude of utility, and hence price.

Consider again the diamonds-and-water paradox. Neither the utility of a diamond or a quantity of water can be higher than the utility of its identical replacement. Hence, provided that water is plentiful and that diamonds are hard to obtain, it follows that water cannot have a high utility, but that diamonds can. Here the paradox is resolved without reference to any "law of diminishing utility". Utility and "scarcity" do not factor into price as two independent elements. Rather possible price is *exclusively* set by utility, and possible utility is limited by performance disutility.

This then provides a second reason why "more costs less". If the earth contains fifty exemplars of a particular commodity in one state, and contains a hundred exemplars of the same commodity in another state, then, all else being equal, the commodities in the second state will tend to be cheaper. What might well happen is that the "performance disutility per unit" of obtaining a specific instance of the commodity will fall. This need not be for any reason deeper than the fact that the commodities in the second state are much more likely to be close to me, i.e. I don't have to walk so far to get them. Here this difference between the two states can be accounted for by saying that the commodity has become less *scarce*, in the sense that a greater amount is available. Or it can equally well be said that the *amount of effort* needed to obtain any given instance of the commodity has fallen. The difference is merely verbal; the important aspect is the ratio of performance disutility to units obtained.

The concept "performance disutility per unit" can account for both the intuitive plausibility of a scarcity theory of value and a labour theory of value. If "more suddenly costs less", then this could well be due to the fact that a fall in performance disutility per unit has lead to a fall in utility, as explained above. If I focus on the "unit" part of performance disutility per unit I will see this as a confirmation of the scarcity theory of value. Simply put, the same effort suddenly produces more of the good. If I focus on the "performance disutility" part, I will see this as a confirmation of a labour theory of value. In other words, it is suddenly easier to obtain a fixed

amount of the good. In this sense the phenomena is equally well "explained" by both ideas.

Another way to state the above is by saying that scarcity is a function of "amount available". Amount is a quantitative notion, but so is "available"; availability is clearly a matter of degree. And this "degree of availability" will have to be a matter of how hard any particular exemplar is to obtain, i.e. "performance disutility per unit". Such an explanation, however, is only partial. For, although performance disutility sets a limit to utility, it is only rarely that it will be equal to it.

The law of diminishing utility is not, then, the only possible explanation for why "more costs less" that was available to the Marginalists. Rather there is another explanation, this one is based on the insight that no commodity can have a utility that is greater than the performance disutility of its identical replacement. This allows for some rough insights into economic processes. Any factor that will systematically change performance disutility will create the possibility that utility might rise, and hence price. The possible factors that can cause such an increase are probably infinite. If something takes less time to make, or suddenly becomes plentiful, or becomes less dangerous, or the social stigma attached to a specific way of obtaining a commodity disappears, or whatever factor is identified that has a systematic influence on performance disutility, then the utility, and price, of such a commodity might well fall[69].

The above has been a rather involved analysis, in order to make a simple point. As regards the measure behaviourally relevant to the determination of prices, *it is incorrect to say that water has high utility*. The above analysis demonstrates why this is not inconsistent with believing water to be a pre-requisite for life. The main point to grasp is that supply-conditions do not enter into price independently of utility, rather they act as a set of constraints on the utility that a commodity can possibly have. This

[69] J B Say was one of the few Classical defenders of a utility theory of value. An inability to understand the above logic, however, seriously seems to have damaged the credibility of his argument. The Classical economists (including Say) agreed that, if the cost of production of a good were to fall, the price would fall. Say also maintained that price is determined by utility. This seemed to lead to a *reductio ad absurdum* of his argument in a dispute with Ricardo, since "Ricardo was able to object that according to Say's treatment, if cost of production fell, utility fell" (O'Brien, 1975: 99). This struck all as absurd, but, as shown above, is actually correct.

insight is the result of realising the full implication of seeing utility as a difference between possible states, and not a quality of an object. It is my contention that, while the Marginalists took a step forward in defeating the idea that utility is a quality of an object, they did not pursue this insight far enough. They still accepted the idea that the total utility of the water in a given exchange is infinite, it was argued above that this rests on a conceptual confusion that they shared with Adam Smith.

There is a sense in which the above analysis is of only historical importance. In the Classical period the diamonds-and-water paradox played an important role in economic theory. It gave the Classical economists an excuse to disregard subjective theories of value in favour of cost-of-production theories of value. The Marginalist solution to the "problem of diamonds-and-water", even if unnecessary, did manage to turn this problem into a non-issue. This paved the way for an agent-relative conception of value, which eventually managed to rid itself of psychological laws like the "law of diminishing utility". In revealed preference theory, which is the modern conception of utility, "utility" is defined in terms of behaviour. Here the *need* for psychological conceptions of utility and performance disutility, as used in the extended discussion of the diamond-and-water paradox above, does not arise. This does, however, imply that something like the diamonds-and-water paradox, which depends on a conception of utility distinct from behaviour, cannot even be stated. Historically it probably does not matter that the way of getting rid of the paradox still contained the seed of the problem that gave rise to it. In this sense the above analysis, if correct, is of interest only to the economic historian[70].

In the context of this thesis, however, it does serve to illustrate a larger issue. This thesis argues for an agent-relative conception of value. As such it is being claimed that agent-independent conceptions of value are fallacious, and lead to unnecessary problems and confusion. This idea is well demonstrated by the history of the

[70] An analysis that has the same consequence with regards to the influence of changes in the difficulty of obtaining a commodity to its price can be found in Hicks' *Value and Capital* (1962). Hicks manages to derive a conclusion consistent with the above *without any reference to psychological utility*, by reformulating statements regarding utility as statements regarding the exchange relation between related goods. The conclusion above is consistent with Hicks' analysis regarding goods that are perfect substitutes (1962: 49). The analysis in terms of psychological, cardinal utility employed above is used, however, since the diamonds and water paradox cannot be stated without reference to psychological utility. Hence any approach that uses behavioural utility avoids the problem, rather than solves (or "dissolves") it.

diamonds-and-water paradox. Here, in the case of Smith, a simple conceptual problem lead to a theory of economic value that was very influential, and is generally conceded to be fallacious. And, even when the problem was resolved by the Marginalists, the solution was still somewhat fortuitous in that it did not get at the root of the problem – the very formulation of the paradox. It also blinded them to a way of relating supply-conditions to utility that did not depend on the "law of diminishing utility", an idea that economists would view with increasing suspicion as time went by[71]. This way of relating supply-conditions to utility can be, as was shown above, a rather useful way of getting a rough, qualitative grasp on the question as to "where prices come from". If the above analysis is correct, then the Marginalists did miss a source of possible insight. The fundamental reason for this is one that I would also blame for some of the confusion regarding value-judgements in philosophy. This is the inability, even when explicitly renouncing an agent-independent conception of value, to fully liberate ourselves from the assumptions and conceptual structures that come with it.

This has been accomplished by the modern economic conception of utility, to which we now turn.

5. The Modern Conception of Utility

5.1 Revealed Preference Theory

As the major accomplishments of Marginalism became evident, economists increasingly turned to its theoretical foundations in order to shore it up against possible attack. Here the "taints" of hedonism and "psychologism" were the biggest issues of concern. One particular problem was the cardinal conception of utility. Here a real breakthrough was made in 1906 when Pareto, using a technique pioneered by Edgeworth, reformulated the results of the original Marginalists without employing a cardinal conception of utility. This did not quite constitute a definitive refutation of the *truth* of cardinality, rather he showed that everything economists need to say about

[71] See Kauder (1965: 135-149)

utility can be stated in terms of ordinal utility, which is the weaker assumption and therefore preferable[72].

The debates concerning the different assumptions of Marginalism need not concern us here.[73] It eventually culminated in the work of Paul Samuelson, who finally stripped the psychological content from the theory of utility in 1938, rendering it purely formal. Today this is known as "revealed preference theory", and works as follows.

Samuelson defines utility purely in terms of the behaviour of agents. If an actor chooses A over B, then she is said to have "revealed a preference for A over B". In this way her behaviour can be used to define the concepts of "preferring" one thing over another or "being indifferent"[74] between two possible options. By employing these concepts the behaviour of an agent can be used to construct an ordinal "preference scale" expressing her preferences with reference to possible combinations ("bundles" in economicspeak) of commodities. Samuelson's great merit here lay in showing that, if such an agent operates with a *consistent* preference scale over time, then this assumption can do all the work that an economist desires from a "theory of value".

This consistency amounts to two requirements: completeness and transitivity. A preference scale is complete if, for any possible choice, an agent can be said to prefer one option to another, or be indifferent between them. The preference scale is transitive if an agent who prefers *a* to *b*, and *b* to *c*, also prefers *a* to *c*.

Hence the fundamental assumption of micro-economics becomes, not the Marginalist "pleasure-seeking", but the consistency of agents. The question now becomes one of

---

[72] This is done by using the idea of an "indifference curve". An indifference curve joins all possible combinations of the two or more goods between which an actor would be indifferent. If this curve has certain properties all the consequences that previously followed from the law of diminishing utility and cardinality can be shown to follow. Hence cardinal utility and the law of diminishing utility no longer needs to serve as a basis for economic theory.

[73] A good summary of the different issues, and how they were discussed, can be found in Kauder (1965: 116-175).

[74] It is not quite self-evident how "indifferent" is supposed to "cash out" behaviourally, unless an actor is suppsed to act like Buridan's ass. Samuelson's work, and the debate as such, took place within a time when there was a general concern with behaviourism so as to render economics "scientific". More than one economist during this time sarcastically asked exactly how long an actor must dither between options before she can be judged "indifferent".

justifying the assumption of consistency. One line to take would simply be to maintain that people simply are, in this sense, consistent. This is an option which few economists would accept, although it has been defended, as will be discussed below. Another line to take would be to treat it as a simplifying assumption that allows for great insight, while not being actually correct. The justification for using it can run as follows. Actors with consistent preferences, and perfect information, would result in markets in equilibrium. Yet it can be conceded that neither preferences, nor information, is ever perfectly consistent or complete. But if the actual operation of the market is studied, the economist does not find total chaos. While the data that is found will never be perfectly regular, it is clear that there is some pattern and order. Hence the two assumptions can be used as idealisations to understand the nature of this order.

These two options can be used by writers who share a common assumption, namely that there is an actual sense in which an individual can be said to "have preferences". Here such a writer would interpret "revealed preference theory" in the following manner: preferences are a reality above and beyond behaviour, yet can only be *known* in terms of behaviour. Such an interpretation takes the phrase "revealing a preference" literally, and not just as a reformulation of behavioural data.

The above way of reading "revealed preference theory" would not amount to the total abandonment of "psychology". Such an interpretation would still be committed to some account of folk psychology, or teleological description, that does not totally "cash out" in terms of behaviour.

The third option is more radical, and can use revealed preference theory" without taking, in any sense, a realist stance concerning teleological description. It utilises the logic of evolutionary processes, and is known as the "money-pump" argument. It works by considering what would happen in a series of exchanges to an agent whose preferences are cyclical, i.e. does not meet the transitivity condition.

Consider a given market, populated by actors who have complete preferences regarding any possible commodity-bundle on such a market. If any actor on such a market has consistent preferences, then the goods of any inconsistent actors will

gradually accrue to her. This is because an actor with inconsistent preferences will, in effect, be willing to trade at erratically shifting exchange-relations that guarantees she will, after a round of exchanges, end up with less of all marketable commodities than she started. If this continues the actor will sooner or later end up with no commodities, and disappear off the market. While the actor can still affect the market by extra-economic means (stealing, begging, etc.), she will have disappeared *qua* economic actor.

Here "evolutionary logic" is at work. Actors are assumed to be consistent, since, if they are not, they won't be around for long enough to seriously upset the model based on assuming they are.

The first point to note about the above is that it is a very elegant method of getting rid of teleological description as an irreducible primary. It cashes out all teleological language into evolutionary stable behaviour; hence actors are treated "as if" they are actualising a set of goals. This is philosophically quite elegant, in that it frees economics from any possible problem that there can be with folk psychology, without seeming to lose anything in the process. Whether it can be applied to a given market hinges on the factors governing the applicability of evolutionary logic in general. It would depend on whether at least one consistent agent exists, whether the actors have a sufficient life-span, and on whether the trades are numerous enough, etc. for the evolutionary effect (elimination of inconsistency) to take effect. Whether this is true will depend on what is taken to constitute the particular "market". In the case of what common usage takes the "economic market" to be it would seem that there is a strong *prima facie* case to be made for this application of evolutionary logic.

The consequence of the money-pump argument that is important for the rest of the discussion is the following. Behaviour can either be of the type that exhibits, in terms of marketable goods, the type of order predicted by economic theory or chaos. The type of behaviour that will result in order must have the following property: *the behaviour of the agents must exhibit a type of consistency that forbids money-pumping in terms of marketable goods*. This will be referred to as "e-consistency".

Behaviour can fail to be e-inconsistent, and still be consistent in a rather obvious sense of the term. Imagine the possibility of a "gift". I wish to give A a certain amount of money, yet I know that A will be too proud to take this. Here I can engage in exactly the type of behaviour that the inconsistent actor who was money-pumped in the example above engaged in. Indirectly I will manage to enrich A, yet he will not be aware of it.

If I do this then, independently of whether it was done intentionally or not, my ability to command marketable commodities has fallen in the same way as has that of the inconsistent actor. If I do this thing a lot, i.e. I wish to give A a lot of goods, or have a large number of people who I wish to enrich, I can eventually end up owning nothing. The market cares not one jot for my motivation, I end up being unable to influence it as economic actor in the same sense as any less altruistic actor. Yet there is an obvious sense in which my preferences can still be consistent. Here I can be said to buy "the knowledge that I have helped A", if this new "commodity" is introduced then my actions can again be consistent with an ordering of preferences that includes this commodity. The ordering of preferences that allows for this sense of consistency I will refer to as "actual consistency", or "a-consistency".

The above point involves some subtlety, and it is vital for the discussion to follow that it is not misunderstood. If one simply defines money-pumping as a process that occurs when an actor is driven off the market *because of inconsistent actual preferences*, then the actor who keeps giving gifts has not been money-pumped. Rather the actor's ability to influence the market has decreased because acquiring the type of commodity ("feeling benevolent") that increases his utility, also decreases his "fitness". Rational choice theory would have no difficulty in pronouncing his actions consistent.

The problem with the above reasoning, however, is that economists require the consistency-condition to have certain behavioural consequences. And a-consistency, i.e. the sense in which rational choice pronounces an actor to be consistent when giving a gift, is *too weak a requirement* to lead to these behavioural consequences. The consistency-condition is used in economics in order to deduce that certain regularities will obtain in the market. But, if any action can be interpreted as the expression of consistent preferences, then the consistency-condition does not *exclude*

any action or occurrence whatsoever. It would be consistent with a market in which all we can find is absolute chaos. This type of hypothetical market would also be consistent with every single actor having wildly *inconsistent* preferences. If even the most chaotic market logically possible would be consistent with both inconsistent preferences *and* consistent preferences, then it follows that the consistency-assumption cannot do what economists need it to do.

Here an extra constraint is needed in order to exclude the expression of consistent preferences, where the behaviour resulting from such consistent preferences could equally welll have been the result of inconsistent preferences. One way to do this would be to state that the actor must be consistent in terms of the commodities *with reference to which the market is defined.* Certain regularities with regard to this market can then be deduced.

Note that even e-consistency would not gurantee that one cannot still disappear from a given market. If I am consistent with reference to the commodities that exist in the market, but keep buying commodities that cannot be traded again, and therefore decrease my market fitness, I can still disappear as actor from such a market. But certain market regularities would still obtain. Therefore e-consistency is here only meant to relate to consistency with reference to commodities that, if I am consistent with reference to them, certain regularities will obtain in the market. E-consistency does not necessarily amount to market-fitness, but a lack of it definitely amounts to non-fitness in a given market.

The above distinction means that an actor can, in terms of a-consistency, be consistently maximising a set of preferences, while also, with regards to e-consistency, be acting inconsistently and being money-pumped. Note that the commodities with reference to which an actor's e-consistency is determined will always be a sub-set of the commodities with reference to which his a-consisteny is determined. What is the to stop an economist from simply including all the commodities necessary for apparent e-*in*consistency to turn into a-consistency?

Two problems emerge here. The type of commodities with reference to which an actor can be said to be consistent might well be commodities like "benevolence",

"prestige", "pleasure", etc. A first objection to including these commodities would be the methodological problem that they might be hard to determine. Methodologically it might be easier to simply assume that actors are a-consistent, where the commodities used to define a-consistency are public objects, and to treat this as a useful simplification.

The second objection strikes to the heart of the money-pump argument. Consider an actor that is pronounced to be consistent with reference to commodities like benevolence and prestige. To an objection that this is not the case, the economist can simply use the money-pump argument, and say that actors are assumed to be so, because, if they are not, they will disappear from the market. This is true if the objects with reference to which one is consistent includes TV-sets, bicycles, and the like. But it does not hold for benevolence or prestige. While we might be able to find some examples where I can be said to be money-pumped out of prestige, it is not at all clear that this is always the case, or that the idea of money-pumping makes any sense whatsoever when applied to commodities like benevolence, sensual pleasure, etc. If I constantly switch between preferring bicycles to TV-sets, then my ability to influence the market by providing bicycles or TV-sets will diminish as I get money-pumped. But if I constantly switch between preferring prestige to sensual pleasure, this need not affect my ability to provide either[75]. Simply put, the fact that I end up with less sensual pleasure than I could have, need not influence my ability to provide it, and thereby influence the market. Hence money-pumps do not apply.

In other words, consistency with regards to the strange commodities needed to turn otherwise inconsistent behaviour into consistent behaviour is not a strong enough constraint to justify anything with regards to economic phenomena. An extra constraint is needed. This could be something to the effect that there is a uniform relation between my position with regards to "strange" and "non-strange" commodities, for example my amount of "prestige" is a uniform function of the non-strange goods I own. But this would be a simplifying assumption with definite

[75] This does not only apply to "strange" commodities, but also services. If the "commodities" on a given market is, for example, "the ability to give legal counsel" or "the ability to play rugby well", etc., then inconsistent preferences need not destroy my ability to influence the market by delivering either. Hence the "money-pump" argument cannot be used to justify the assumption of stable preferences with regards to these commodities.

counter-examples, and no improvement on simply talking in terms of e-consistency, and skipping "strange" commodities altogether.

The fact that a-consistency allows for money-pumping *in terms of a given market* means that a-consistency does not necessarily amount to e-consistency. It also seems to mean that, since what is needed for the application of economic theory to a given market is e-consistency, showing that apparently strange behaviour is actually a-consistent does nothing to justify the application of economic theorems to explain the actions of a-consistent actors. This can only be done once it has been shown that the specific case of a-consistency implies e-consistency.

The discussion above, and the distinction between e-consistency and a-consistency, will be of some help in understanding the claims of Gary Becker, an economist who, as a matter of principle, *never* takes recourse to inconsistent preferences in order to explain human behaviour.

5.2 Becker's stable preferences

The assumption of e-consistency amounts to a constraint on behaviour. Yet micro-economists are often gifted at finding reason and consistency in behaviour that violates e-consistency in the most flagrant manner. Here no-one stands taller than Becker, where most would cry "inconsistent" or "irrational" he is always prepared to defend the consistency of people's actions. The difficulty with this is that it becomes unclear whether Becker's sense of consistency still places any constraint on human action. If Becker's sense of "consistent" can be applied to any set of behavioural data, then it does not serve any obvious purpose. It seemingly cannot be used to justify economic theorems, since, as explained above, economic theorems do, at least sometimes, tell us that some things will not happen.

In his justly famous essay, *De gustibus non est disputandum* (with George Stigler), Becker claims that the preferences of people do not change[76]. He also doubts whether different people have different preferences, this doubt seems to extend across different

[76] "[O]ne does not argue over tastes for the same reason that one does not argue over the Rocky Mountains – both are there, will be there next year, too, and are the same to all men" (1977: 76).

cultures and different times. He chastises economists who, when faced with a seeming change in preferences, accept this as the explanation of a change in behaviour. Becker treats it as a fundamental methodological commitment that this is never the case. He states that to accept such a preference-shift is just to admit ignorance of the true consistency underlying the apparent inconsistency (1977: 76).

Becker is normally read in two different ways: either as claiming that allowing a shift in preferences is bad methodology since it gives up on the very idea of explaining a change in behaviour, or as actually insisting that preferences are stable, period. I think that, once one gets used to the seeming bizarreness of the idea that preferences actually are stable, and not even varying between different people, it becomes reasonably clear that Becker wishes to make both claims[77]. This is certainly the more interesting of the two interpretations and, regardless of whether Becker consistently maintains it, an idea that can be fruitfully discussed within the context of this thesis.

Becker's work has, predictably, been controversial. This is due, both to the above claims, and to the fact that he does not stay within the realm traditionally reserved fro "economics". By "economics" Becker means an approach, rather than a subject-matter. The essence of this "economic approach" is the assumption of maximising behaviour and stable preferences under conditions of scarcity (1976: 5). Becker uses these assumptions to investigate areas of human behaviour not normally discussed by economists. He states that the economic approach can be applied to any human action whatsoever[78]. Examples of topics that he has treated in this manner include discrimination (titled: "Price and Prejudice"), criminal punishment, and, notoriously, fertility and marriage[79].

---

[77] He explicitly states that his assertion regarding consistent preferences is "an assertion about the world", but does say that the choice between viewing preferences as mutable or not must be based on the fruitfulness of the results achieved by using these assumptions (Stigler and Becker, 1977: 76).

[78] "[A]ll human behaviour can be viewed as involving participants who maximise their utility from a stable set of preferences and accumulate an optimal amount of information and input in a variety of markets" (1976: 14).

[79] These topics, and others are covered in his 1976 collection *The Economic Approach to Human Behaviour*. The essays regarding marriage and fertility are written in terms of categories like "the marriage market", "the optimal sorting of mates", and "the quality of children". These types of expressions are, unfortunately, often enough to stop some writers from even considering what he has to say.

The resonance of Becker's work with this thesis is obvious. He views all human behaviour as the maximisation of utility. This thesis must ultimately agree with this statement, but add that "pleasure/goodness" is the only "true utility".

It is not, however, this side of Becker's work which will be explored here. Rather we will try to make some sense of his claim that preferences are as constant as the Rocky Mountains (Stigler & Becker, 1977: 76). If Becker is right in this, then surely he must mean that people are a-consistent. For surely the rather obvious example of a "gift" used above would make nonsense of all human behaviour being e-consistent.

Becker does refer to what was defined above as a-consistency. The exception to e-consistency because of the gift was turned into a-consistency in the example used earlier by introducing a nonmarket commodity, "the knowledge that I have helped A". Becker achieves a-consistency by continually introducing these type of commodities when faced with an apparent shift in preferences. Indeed, when it comes to "commodities", Becker is almost scornful of "ordinary commodities":

> That preferences are assumed to be stable do not refer to market goods and services, like oranges, automobiles, ..[but to] fundamental aspects of life, such as health, prestige, sensual pleasure, benevolence, or envy, that do not always bear a stable relation to market goods and services (1976: 5).

5.3 Relevance of Becker

5.3.1 Intentionality

An evaluation of Becker's work falls outside this thesis (and is beyond my competence). Below an attempt will be made to make it appear less strange, and to relate e-consistency to a-consistency. One initial point should, however, be clear. Becker is fully committed to teleological description, he cannot merely be talking about actors "as if" they pursue goals. As such the flight from "psychology" and teleological description that seems to reach its peak in the work in Samuelson has now

come full circle. Becker views an agent pretty much as Bentham would have viewed him, except that he uses actual preferences as a base rather than pleasure and pain[80].

If Becker does not think that there is an actual sense that an actor can "have a preference", above and beyond mere behaviour, then all his arguments would be circular. For any "seeming" problem can be made to fit without even trying to be clever about it. No two choices will ever, if we are prepared to go into minute detail, have the exact same content. The most minute detail in the specific content of the choice of an actor can merely be stated to constitute to a "new commodity", and victory declared.

I wish to make it clear that the above concern is related to more than the truism that an infinite number of (increasingly bizarre) utility-functions can be constructed that would be consistent with any specific set of behavioural data. If Becker is not claiming that there is a matter of fact, over and above behaviour, in virtue of which an actor can be said to be consistent or inconsistent, then the controversy surrounding his work makes little sense. For if he is not assuming that preferences are irreducible to behaviour, then he is only claiming that, for any behavioural data, a reasonable simple, analytically useful and consistent utility-function can be constructed. This is a much weaker claim than the claim that there is a sense in which people actually are consistent, and it is hard to see that this claim could have generated the same controversy. It also seems clear that Becker does not understand his own work in this way. When he compares preferences to the Rocky Mountains (1977: 76), he is presumably not claiming that the Rocky Mountains are a useful theoretical construct that can be used to explain certain data. Hence Becker's work only makes sense if he is interpreted as relying on the idea of "truth-conditions" for his claims, above and beyond mere behaviour.

5.3.2 A-consistency and e-consistency

There need not be anything but a verbal disagreement between someone who asserts behaviour to be inconsistent, and Becker's insistence that this is never the case. When

[80] Becker cites Bentham with approval (1976: 8).

dealing with e-consistency, we are dealing with a specific market. The assumption of consistency within this market means that the preferences of actors within this market can only be given in terms of the things allowed to count as "commodities" within this market. There can be good reason for restricting the definition of "market", the best is surely that the definition of a market and the actors in it allows for the assumption of stable preferences based on the money-pump argument. In other words, the market is of such a kind that evolutionary logic can take effect. Such an approach can, surely, be useful.

If an action is pronounced inconsistent in such a market (in terms of the behavioural standard), then, as explained above, the analyst need not deny that there is another sense in which the behaviour is consistent. It might have been a case of "surreptitious gift-giving", or some other "strange" commodity being acquired. A Becker-style analysis would make the action consistent by adding a commodity, the "joys of gift-giving" or whatnot. This would be a complementary analysis, not something in conflict with the first analysis. These analyses must ultimately be judged by the fruitfulness of their results, not any dogmatism about the stability of preferences.

There is, however, a downside to expanding the definition of "market". If a market is defined so as to include pretty much *anything* as either a possible "good" or a possible "price / shadow price", the money-pump argument no longer works. The logic of money-pumps can destroy someone in terms of e-consistency, and yet this person remain a player on Becker's market. As long as any action of mine can influence the utility of another actor - and a situation where this is not the case is almost unintelligible - I can use this as leverage. Hence I can still "trade", in the widest possible sense.

Again the upshot of this, in the context of this thesis, is that teleological description can then no longer reduce to "as if"-talk. Hence the folk psychology that was so suspect in economics earlier in the century reappears.

E-consistency is always relative to a specific definition of a market, where this definition excludes some elements of reality from being either commodities or influencing the budget-constraint. A-consistency refers to the actual preferences of an

actor. If there is such a thing we certainly have no grounds for the *a priori* exclusion of anything from counting as a commodity or price in terms of it. This also implies that showing a set of actions to be a-consistent does justify a treatment of such an action by using economic theorems, etc. It does not justify using the theorems for markets that have been specified in terms which exclude the use of the goods needed for a-consistency. But it does justify the use of the theorems to study the same actions, but using a different definition of market that includes these commodities.

5.3.3 Are Preferences Stable?[81]

5.3.3.1 Hedonism

Can preferences be said to be stable? In terms of the position of this thesis (and any hedonist position), the answer is yes. Hedonism basically amounts to saying that, ultimately, pleasure is the "real good" which is always bought and sold. More pleasure is always preferred to less pleasure (or: better is preferred to worse) and hence preferences are transitive.

The more interesting question concerns the relation between pleasure and all the commodity-bundles in the world. If a "state" is all the true facts at a particular time, and this is taken to be the ultimate commodity-bundle, then the question can be asked whether these commodity-bundles are transitively ranked. This would seem to depend on a claim of supervenience. If the pleasure obtained in a state is a strict consequence of the state, and the state alone, then the relative pleasure in two states cannot change without a change in the states themselves. This change in the states would then be the explanation of the changes in pleasure. If such a supervenience-relation holds, then hedonism implies that all possible states are transitively ranked, and that this ranking is fixed[82].

---

[81] A slightly different way of going about this would be to simply define an "agent" in terms of stable preferences. This is proposed in Ross (2002). Such a conventionalist move would lead to rephrasing the above question as the question "are people agents?". Ross answers this question in the negative, and states that people are, at best "*sequences* of economic agents" (2002: 101).

[82] A strict claim of supervenience would even seem to erase the problem concerning intersubjective utility-comparisons if all phenomenal qualities are taken as strictly supervenient on the non-phenomenal. The very notion of a "subject" disappears, the non-phenomenal difference between two people simply simply becomes different facts of different states. The non-phenomenal is always commensurate, if the phenomenal supervenes upon the non-phenomenal then no "added element" can

In the above two states preferences, ultimately expressed in terms of states that include all true facts, are stable. Can the same be said about Becker's "stable preferences"?

5.3.3.2 Becker

Presumably Becker would not wish for the *a priori* exclusion of any possible elements of reality as a possible commodity or shadow-price. If all these possible elements of reality are again spoken of as states, then it is apparent that a claim of supervenience would imply again that he is correct. I'll define "wantedness" to be the "whatever" that causes one state to be chosen over another, regardless of whether this is one element (hedonism) or a complicated set of interrelated factors. If "wantedness" is supervenient on the rest of a state, then this amounts to saying that the wantedness of a state is fixed, unless some other element of reality changes. This change then serves as the explanation of the change in wantedness. If this is correct, then there is a fixed ordering of possible states.

The above would seem to suggest one way of interpreting Becker which can remove quite a bit of confusion, especially where charges of circularity, etc. are involved. Rather than treating the central distinction as one between stable preferences and shifting prices, his statements can be "translated" into statements concerning stable "possible states" and shifting "reachable states".

Consider the following: An economist has defined a market in terms of a given set of commodities, and has assumed stable preferences in terms of these commodities. This has been justified by using a money-pump argument. Let two of the commodities be "hours spent mountain-climbing" and "hours spent bike-riding". If someone suddenly reveals a preference that contradicts her earlier preference, and the budget-constraint (defined in terms of the commodities that were specified to constitute the market) hasn't shifted, then the economist will pronounce her inconsistent. If Becker should

---

be introduced to suddenly destroy this commensurability. If this is correct then all possible states of all possible people form one transitive ranking.

dispute this and say that prices have, somehow shifted, this need not be any more than a verbal disagreement. The economist might admit that there might be a "shadow-price", but that it is not one of the things which can be a price in his analysis, and that he has good reason for having excluded it from his definition of "price".

Assume that the person in the above example (call her Active) has suddenly gotten a bad case of vertigo. Now Becker might simply say that mountain-climbing has acquired a shadow-price – the fear experienced while climbing. Now a bigger, more complex market can again be defined, in terms of which she has always been consistent.

The above has been cursorily explained before, but the example was created to answer a question, namely: does Becker only have a verbal disagreements with those who have made him controversial? Is there a matter of fact at stake here?

I would argue that there is, and that this can best be shown by drawing a distinction between *possible* (logically possible) and *reachable* (physically possible for the actor) states. In the above example, Becker can say that Active still prefers the state in which she climbs mountains and has no fear to the state in which she rides a bicycle. But the vertigo means that "climbing-without-fear" is no longer a reachable state for her. Rather the state in which she rides a bike is now preferable to the state in which she climbs a mountain while afraid.

In this way Becker's position can become the claim that all possible states are transitively ranked, all that ever changes are the states that it is possible to reach. An "explanation" of a "shift in preferences" now becomes an explanation of why a formerly reachable state is no longer reachable, or *vice versa*. A "commodity" can now be defined as any element of reality that enables one to reach a higher state than without it. A "price" or "shadow-price" can be defined as any element of a state that impedes one from reaching a wanted state.

In this version there is, if one accepts intentional description, a matter of fact at stake. This is the claim that there is a fixed ranking of possible states. This saves Becker's

claim from meaninglessness or circularity. I would also contend that it makes it a good deal less strange than the formulation in terms of "prices" and "preferences"[83].

There is a methodological problem in that it seems difficult to determine what would decide the issue regarding whether possible states are transitively ranked. There is the related problem that it seems hard to determine, even if they are so ranked, whether a given explanation of behavioural changes is true or false. A state can become reachable or unreachable due to a change in brain chemistry or matters similarly hard to determine. Hence assertions regarding specific behaviour might be, practically speaking, non-falsifiable. This problem will not be considered here. For now it will have to suffice to have shown that what Becker is asserting is not a mere tautology.

5.4 Relevance of the Discussion of Revealed Preference to this Thesis

The above discussion, despite its twists and turns, was guided by a single question: can the economic conception of value get rid of intentionality? The answer seems to be that it can, but only partly. It can inasmuch as the teleological idiom can get cashed out in terms of money-pump arguments. It was shown above that such "cashing out" is limited by two factors. The first is that the specific market under discussion might not have the particular qualities that enable evolutionary logic to take sufficient effect. The second limit is, in a sense, a special case of the first. This is were the definition of "commodity" becomes so wide that it starts to include commodities that fall outside the logic of "money-pumps". This would be commodities like Becker's "prestige", or "sensual pleasure", etc[84].

Hence economics can avoid intentionality, but not if they wish to say all the types of things that economists wish to say. The occupation with a behaviourist standard and extensional description does not seem to be taken all that seriously anymore. In this sense it can be said that, while economics managed to, in principle, strip away the psychology at one point, they didn't wait all that long before coming back to it. It

[83] On this formulation it must surely be apparent that Becker's refusal to allow for a shifting of preferences amounts to no more than the assertion that explanation of a change in behaviour is always possible. His assertion that preferences are stable is basically the assertion that there is a type of "principle of sufficient reason" for economics.

should be noted, however, that the psychology it came back to is of a considerably minimalist type. Instead of dubious stuff like cardinal utility or the law of diminishing utility, it seem only committed to two claims. The first is that teleological description is, in some sense, okay to treat as an irreducible datum. The second is that preferences are stable, or at least stable enough to form useful models.

The issue of intentionality has been extensively discussed in philosophy. Although it is by no means clear that realism about intentionality is wrong, it is clear that such a position is not self-evidently sound. But it would surely be churlish and "overly philosophical" for a materialistically inclined philosopher to chide economists for their reliance on this idea[85]. Within the context of this thesis, however, such reliance takes on added relevance. If economists need intentional description, then, I would argue, hedonism is back.

The economist, if he is a realist about preferences, is basically taking the position that was referred to in the first chapter as "treating wants or needs as fundamental data"[86]. It was argued that the category "want" cannot be logically fundamental, rather it is derived from the category "good". If this argument is correct, then, in committing to "wants", an economist is committed to "good". It was also argued tat this "good" is what we often refer to as "pleasure". If this is correct then the economist is committed to hedonism.

I do not wish the above claim to be misunderstood. It would be silly (or at least unnecessary) for any economist to call himself a hedonist (*qua* economist) based on the above concerns. The arguments advanced in chapter one, while I would contend that they deserve consideration, are not definitive proofs. They depend on tricky notions like "primary", "qualia", intentionality", etc. that are far from rock-solid. It would also serve no purpose for an economist to become a hedonist. Whether "wants"

---

[84] Another way of formulating this would be to say that the distinction between "actor" and "economic actor" gets obliterated.

[85] It would seem almost akin to asking economists to give up the perfectly serviceable distinction between a definition and a statement because of Quine's attack on analyticity. This has been done. See Hutchison (2000: 204-207) for an argument against this type of philosophical intrusion.

[86] "Money-pumps" are, off course, not the only way to "cash out" teleology into something else. One could also talk about taking the "intentional stance" toward a system, *a la* Dennet. This option is not explored in this chapter, both because irreducibility of teleology is an assumption of this thesis and because this is not how Becker and company understand their own statements.

are treated as primary or derived would have no impact on the theorems of economics, so accepting only the weaker assumption would seem to be the merest common sense. Hence I would say that, while economists might well be committed to hedonism if the arguments regarding the unintelligibility of "want" as primary is correct, they won't in any obvious sense be served by taking cognisance of this possibility[87].

This does not quite apply to any user of the modern economic conception of value. For the position in this thesis does have one major advantage – preferences can be judged. Preferences, or, the specific things that are wanted, are not irreducible primaries for the hedonist. Rather they are the result of some conception regarding what type of things will lead to pleasure. As such the question whether wanting a specific thing will lead to pleasure concerns a matter of fact; preference are "cognitive". This has one very important consequence. The economic conception of value does not allow for a person to "act against his own best interests", except where actions are based on mistaken beliefs regarding the world. But if someone was to, for example, systematically "reveal a preference" for cutting themselves with knives the economist cannot declare this bad or even irrational. A hedonist has an extra option, since a person's preferences are themselves "cognitive". The hedonist can, if it is indeed true, declare the person to be acting against his own interests.

According to the way academia tends to split things up we would not think this type of judgement to be the job of an economist. Hence the assertion that the economist need not be hedonist can be saved by a conventionalist move. This would be to say that, if an economist does pronounce preferences irrational for the above reason, she is not acting as an economist. This type of thing is largely arbitrary, what is important however, is that the hedonist position would avoid the possible quietist consequences that would come from having only the modern economic conception of value.

---

[87] The economist Richard Layard has recently come out in favour of hedonism in economics. In a set of papers given at LSE he portrays pleasure as a measurable cardinal quantity in the brain, and actors as pleasure-maximisers. This is done in order to make the claim that economic growth, while a laudable goal, does not necessarily result in increased happiness (Layard, 2003).

6. Conclusion

This chapter has attempted to give an overview of the main historical conceptions of value, and to relate these to the position developed in this thesis. It was firstly shown that the notion of value can be related to economic phenomena in a rather simple way, and that this way is basically similar to the one employed by the Marginalists. It was argued that both the Marginalists and their predecessors suffered from some conceptual confusion regarding the conceptualisation of value that lead to unnecessary problems. The greatest of these were the so-called "diamonds-and-water paradox".

It was argued that the major move in economic history, with regards to the theory of value, was a move away from an agent-independent to an agent-relative conception of value. Hedonism was the historical catalyst for this change, yet it ultimately served as a Wittgensteinian ladder that was cast off at no cost. This "casting off" even went so far as to be able to cash out all intentional talk into evolutionary terms. This type argument is not, however, always used by economists to justify the assumption of stable preferences[88]. Economists still need intentional talk for certain types of explanations and analyses. If the argument in this thesis is correct, then this intentional talk brings economics straight back to hedonism.

I wish to note a final point before proceeding to the next topic of discussion. This thesis is trying to argue for a meta-ethical position that has, as one implication, hedonism. Hence it needs to, at some point, attempt to argue that all human actions can, in some faint sense, be seen as expressions of egoism. This is not a new position in philosophy, but the task of making such an argument is quite daunting. Here I think it justified to claim that the history of economic analysis does give such a claim an element of plausibility.

Economic interaction is perhaps the area of human behaviour that we are most comfortable with describing in selfish terms. Economists who definitely viewed man

[88] The money-pump argument does not dominate discussions regarding the stability of preferences. It is unclear exactly how large a role economists view it as playing, some standard economic reference works do not even mention "money-pumps" (or "dutch books", etc).

as more than selfish (e.g. Menger) had few qualms about studying man's least edifying activity in terms of self-interest, but maintaining that outside the economic market he is frequently altruistic in a qualitatively different sense. As such it is natural that the assumption of maximising behaviour[89] first took its firmest root in economic theory. The concepts developed, however, and the approach as such did not stay confined to economics for long. "Public choice" theory, "social action" theory, and the myriad uses that have been found for various decision theoretical concepts testify to this fact. These uses all, in some sense, are the result of considering what happens when two self-interested individuals engage in economic trade.

This is not to say that the above theories have always swept all before them, nor to deny that there are areas which are at present intractable to such analysis. I would, however, contend that the refusal of "maximising behaviour" to stay within the realm of "economics", narrowly construed, is something from which the defender of egoism can take heart. Since Becker is here the undoubted hero, this chapter will end with his summary, in economicspeak, of part of what is being defended here.

> The heart of my argument is that human behaviour is not compartmentalized, sometimes based on maximising, sometimes not, sometimes motivated by stable preferences, sometimes by volatile ones, sometimes resulting in an optimal accumulation of information, sometimes not. Rather, all human behaviour can be viewed as involving participants who maximise their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs in a variety of markets (1976: 14).

With that said, I now wish to claim that there is an important sense in which people are not maximisers. This will form the topic of the next chapter.

---

[89] It might be doubted whether "maximising behaviour" really amounts to egoism. This would largely be a matter of defining "egoism", but I would contend that there is an important sense in which it does.

Maximising behaviour has to, in some sense, relate all actions to the interests of the actor, even if these interests can include the interest of others as object.

Chapter 4: Towards Ethics: from Calculations to Rules

1. Introduction

Chapter 1 and 2 explained and defended the idea of cognitive, ethical, egoist hedonism. It was argued that this doctrine presents an elegant solution to epistemological problems regarding value, and that it can overcome the traditional objections to the idea of a cognitive ethics. It was also explained that ethical hedonism results in a variant of psychological hedonism. Chapter 3 related this conception of human action to economic phenomena. It also attempted to show that economics has historically been well-served when it has employed this assumption (or parts thereof) to explain human behaviour.

There are two unfortunate (or counter-intuitive) implications of the theory of cognitive ethics that is defended in this thesis. The first is that ethics is hedonist. Hedonism has a long philosophical tradition, and though it is unpopular nowadays[90], is not totally unrespectable. Yet I would guess that, if you asked philosophers what ethical theory they would *like* to be valid, i.e. if their wish could magically change the metaphysics of the universe, I do not think you would find many hedonists. Hedonism as an ethical criterion isn't quite as inspiring as the categorical imperative or "universal harmony" or the "inevitable dictatorship of the proletariat", etc. If it were up to me I would choose a criterion that is less worldly and transient than hedonism. Yet, if ethics is cognitive, as is being argued here, it isn't up to me in the same sense that the law of gravity isn't up to me.

The other unfortunate (or counter-intuitive) implication, one that is infinitely worse, is egoism. If one were to ask most people what they consider the very *antithesis* of morality or "ethical action" to be, the answer would probably be egoism. Or, rather, "selfishness". Ethics is sometimes even loosely defined in terms of its supposed opposite - egoism - so that egoist action is "unethical" by definition. If the definition of ethics used in chapter 1 is employed, and ethics is defined as an attempt to answer

[90] A survey among contemporary philosophers will not turn up a lot of hedonists. It is probably fair to say that a lot of people who would have been hedonist-utilitarians in an earlier age are preference-utilitarians today. Here the way economic theory has developed from Marginalism to modern micro-economics has clearly had a large influence.

the question "How should I/we act?", then this rather seems like an attempt to do by definition what needs to be done by logic. The egoist implications of the theory under discussion are the direct consequence of classifying "the good" under *qualia*. This idea, if there is something to it, is again as unalterable as the law of gravity.

The news, however, isn't all bad. The next three chapters will attempt to argue that, in fact, the ethical consequences of the meta-ethical view adopted in this thesis might not be nearly as unfortunate as one might suppose. Rather there is reason to believe that our intuitions regarding the forms that egoist action will take are radically untrustworthy.

In order to construct such an argument the question that needs to be asked is the following: how should a hedonist egoist act in order to maximise value? This question is surprisingly complex. The next three chapters will only scratch the surface in its attempt to answer this question. This chapter will attempt to argue for a first, and vital link between egoist hedonism and what is commonly understood as "moral action". It will be argued that an egoist hedonist should, in an important sense, be a rule-following creature rather than a calculating creature. Chapters 5 and 6 will then concern some of the rules that such a creature should follow.

The first step in this argument is to discard the simplifying assumption used in the previous chapter. This assumption was that people can directly calculate the value of the different options open to them, and then choose the best one. This entire chapter will be an attempt to explain why no egoist hedonist should try to do this.

Note that this chapter will, in the first instance, be concerned with all human action, not just those we consider "ethical". "Ethical"[91] action is a subset of human action as such, and the former can best be illuminated by considering the latter.

---

[91] On the definition of ethics used above, *any* attempt to direct any human action counts as "ethics". This is congruent with the main claims of this thesis which, at base, only recognise one standard for all human action. Sometimes, however, I will use the word "ethics" in a narrower sense. This is a sense in which advice on how to eat a sandwich does not count as "ethical", but advice on whether to save a drowning man does. I trust that context will clarify usage sufficiently so as to prevent confusion.

Also note that, if the reader did not find the arguments in chapters 1 and 2 convincing, that nothing of substance in this chapter depends on the arguments made there. If those arguments are correct, then the next logical step is to determine what an egoist hedonist should do. Hence this question will be considered in the present chapter. But an egoist hedonist is just one *species* of utility maximiser[92]. And the topic of this chapter is substantially the same as the question "how should a utility maximiser act?". This is an interesting question in-itself, regardless of the work it is doing in the context of the thesis. The general points in this chapter - regarding the relation between rational choice theory and bounded rationality - has no need of a hedonist, egoist base.

2. Definitions

The mind is an information-processing mechanism that allows a utility-maximiser to attempt to maximise utility. The decision-rules according to which information is processed by the mind can be distinguished based on whether all relevant and available information is used, or whether some relevant information is ignored. I will refer to a decision-rule that uses all information as *calculation*, and one that does not as *rule-following*. In this chapter it will be argued that man is a rule-follower in a sense more basic than he is a calculator.

The above definitions need to be explained so as to prevent confusion:

- Calculation: To calculate is to follow a decision-rule that uses all available information relevant to a specific problem.

The following is the type of thing that counts as a calculation: Take it as a given that I know I want the cheapest meal on the menu at a restaurant. If I look at every price on the menu, identify the cheapest meal and choose it, then I have calculated. Hence I have calculated *relative* to the finite set of options on the menu and the standard of price.

---

[92] This seems to be why Broome states that "expected utility theory [i.e. rational choice] describes the structure of good" (1994: 138).

- Rule-following: A rule is here used to denote any decision-rule that does not use all the available information relevant to a specific problem.

Consider, again, the situation of being in a restaurant if one has very little money. Imagine the menu is extremely long, so that going through the whole menu in order to determine the cheapest item wastes too much time to be practical. I am trying to eat as cheaply as possible, but, since the menu is too long to read, I follow the following decision-rule: read from the beginning of the menu, and choose the first dish that costs less than R30.

Note that, according to the above definitions, the above rule only counts as a rule because of the specified goal of action. If the goal of action had been to order the first item on the menu under R30 in price, rather than to eat as cheaply as possible, then I would have been calculating when I looked at the menu in order to determine the first item under R30 in price. Hence the above action can be analysed as follows. In order to eat as cheaply as possible, I followed the rule: "choose the first option under R30 in price". This rule can now be treated as a sub-goal. In order to achieve this sub-goal, I calculated by looking at the menu and choosing the first option that satisfied this criterion.

If information-processing can be said to be a calculation with reference to a sub-goal, I will refer to this as *relative calculation*. If information-processing can be said to be calculation with reference to an ultimate goal, i.e. with reference to maximising utility as such, this will be referred to as *absolute calculation*.

A first implication of the above definitions is that a being with a constrained decision-making capacity can never be an absolute calculator. Man's decision-making capacity is constrained by the fact that decision-making takes time, and that time is a finite resource. Furthermore, the possible options at any given time in life is infinite. Here a typical "economic" problem, namely how to optimally distribute finite computational capacity over infinite possibilities, arises. This thesis will argue that an examination of this problem can go a long way towards giving an account of what we call "moral action". For now let it be noted that the possibility of calculation concerning everything I do while doing it is almost unintelligible. For a start, a creature that could

do this would need *instantaneous* information-processing capacity. In other words, such a creature would have to be able to perform an infinite amount of calculations, without any of them even taking *any* time. Hence an absolute calculator, as defined above, seems a mere logical possibility.

This leads to a crucial issue with regards to the relative status of calculation and rule-following. Rule-following ignores certain information relevant to a specific problem, hence it is computationally cheaper. Since information-processing capacity is a finite resource, it is preferable to follow a rule than to calculate, *all else being equal*. Often, off course, all else will not be equal. Rather the following of a rule is likely to lead to a loss of accuracy that negates the saving in computational capacity. This chapter will consider the nature of this trade-off in detail.

Note that no assumption is being made in any of the above definitions about whether a specific act of calculation or rule-following occurs consciously or unconsciously. Whether a specific act of information-processing occurred consciously or not is irrelevant to the main points being made in this chapter, except where explicitly stated otherwise.

With these categories defined the question as to how a hedonist egoist should act so as to maximise value can now be investigated.

3. Three Sources of Rule-following

3.1 Ignorance as Source of Rule-following

It will be argued that there are four main sources of rule-following, i.e. four conditions that, when they obtain, imply that rule-following will be preferable to calculation. The first three can be understood as operating on the same logic, and can be illustrated in the following manner.

Imagine a benevolent being has twice hidden gold coins, in a hundred clearly marked holes each time, on two different planets. The being had used a pseudo-random number-generator to decide whether a given hole should contain a coin. The standard

used was of such a nature that, for any particular hole, there is a 99% chance that it will contain a coin. Imagine these coins are valuable, and that digging them up takes no great effort.

Now imagine that one being populates each of these planets. The first is omniscient and has instantaneous decision capacity; i.e. it takes the being no time to make a decision[93]. This being we shall call Perfect.

The second being has roughly the amount of information and computational capacity that would usually be expected of an average human being. Let's call this being Normal. Assume that the mathematics of pseudo-randomness is something that is simply beyond Normal's ability to understand.

How will Perfect act? Perfect, being omniscient, will dig up only holes with coins in them. Holes that do not contain coins, if any, will be left untouched. Being omniscient she will be able to *calculate* what "decisions" the use of the pseudo-random number-generator would have resulted in. Perfect's actions in digging up these holes can be explained solely with reference to the notion of calculation. Since it is intrinsic to a rule that some potentially relevant factor is not considered, Perfect can in no sense be said to have followed a rule.

Normal will dig up all hundred holes. This example was constructed in such a way that the chance of digging up a hole unnecessarily is far outweighed by the chance of a benefit derived from finding a coin. Normal will know that at least one hole is probably empty, *yet this should not change his actions one bit*. His knowledge that at least one hole is probably empty is useless to him.

Did Normal's actions constitute following a rule or calculating? The answer is that Normal's actions in digging up a hundred holes could have been the result of both. It is possible that, *vis-à-vis* the action of digging up a hole, and in terms of finding a coin as a standard, Normal calculated afresh each time he saw a hole. In other words

[93] It is questionable whether this possibility is really intelligible. For the purposes of the above example it wouldn't really matter if the being was alternatively thought to have the fastest computational capacity conceivable.

he saw the holes, each time considered the possibility that it might contain a coin or not, and each time decided to dig. But it is also possible that Normal could have followed a rule. He could have calculated right at the beginning and, *vis-à-vis* the possibility of it containing a coin or not and with the desirability of finding a coin as choice-criterion, decided to dig whenever he finds a hole. Hence, when Normal saw his second hole, he decided to dig right after identifying it without running through the calculation again.

Hence Normal's action can be the result of a hundred acts of calculation, or one calculation and a series of acts best described as rule-following. The above example might seem somewhat fanciful, but it does illustrate a first important point. This is that ignorance can lead to *uniformity* of action. Normal's ignorance of the workings of the pseudo-random number-generator's workings results in a situation where he has to dig up each hole. This was not the case with Perfect; she only dug up the holes with coins in them.

The difference between calculating and rule-following lies in the fact that, in rule-following, a possibly relevant factor is ignored. Normal can ignore the possibility that any particular hole might not have a coin, considering this possibility will not change his actions. Hence ignorance can result in a situation where calculation and rule-following have the same result. This leads to the conclusion that ignorance (or: informational constraints) is the first potential source of rule-following.

3.2 The Cost of Information as Source of Rule-following

In the above example Normal simply cannot determine which holes have coins in them, as the mathematics of the pseudo-random number-generator is beyond him. Now imagine that the situation changes. Every time a hole is dug, a sign is placed a hundred meters from the hole. This sign indicates whether the hole contains a coin or not. Now Normal can walk to the sign, read it, and decide to dig based on what the sign says.

Yet it would not be in Normal's interest to do so. The cost of the time spent walking might well outweigh any gain in accuracy. The only saving that such a course of

action would result in is to prevent him from occasionally digging up a hole unnecessarily. The example was constructed in such a way that this effort is minimal.

Hence Normal can continue to ignore a potentially relevant factor when it comes to digging up holes. He could still follow a rule rather than calculate, and with no ill effect. This time, however, he is not ignoring a possibly relevant factor because it is beyond his ability to know it. This time he is ignoring it because the *cost* of determining it, i.e. the time invested in walking, is too high. This leads to the second important point that can be illustrated with the above story. This is that the cost of information can be a source of rule-following.

3.3 Computational Cost as Source of Rule-following

The third source of rule-following is really a subspecies of the second. This is where the cost of information is again too high to justify calculation, but for a very specific reason. This is that the high cost is specifically due to the cost of computation.

To illustrate, consider the following. Normal encounters all of the holes, but they are all encountered on the same day. Normal knows that he has to take the coins on the given day. The next day they will be gone. Also assume that the coins are buried a bit deeper this time, so that the cost of digging rises.

Next imagine that the cost of determining which holes have coins has been lowered. The being that placed the coins there had left markings to indicate which holes have coins. These markings are in a code that can easily be deciphered by Normal. Imagine that the relative pay-offs of Normal's options are such that he is best served by quickly figuring out the code each time he finds a hole, rather than digging. Hence Normal is now not digging up any holes unnecessarily.

But also imagine that Normal has some project that he is working on, and that he needs to come up with a fairly detailed plan of action regarding this project. He can do this in his head while walking, but needs all his concentration in order to keep things straight in his head. Stopping to figure out the codes in order to determine whether to dig breaks his concentration repeatedly. Not only the time incurred in

figuring out the codes is responsible for this, but also the fact that it takes a while to recapture the "mental grip" he had on his plan before his concentration was broken.

Normal can abandon the coins, and simply concentrate on his plan. Or he can abandon the plan, and collect the coins. But his computational capacity is of such a type that he cannot figure out his plan, and acquire the coins by cracking codes.

Fortunately Normal has another option available to him. He can simply follow the rule of digging whenever he sees the hole, and dig up all the holes. Assuming that he can concentrate while engaged in physical labour, his mind is now free to concentrate on the plan[94].

The above example illustrates an important point with reference to computational power. A rise in the computational cost, namely that these computations were breaking his concentration, can again lead to uniformity of behaviour. Hence computational costs are another possible source of rule-following.

There is also a more subtle matter with regards to computational costs, and informational costs in general, which is quite important. Imagine the cost of determining whether a given hole contains a coin is less than the (decontextualised) cost of digging a hole. It might still be optimal to forego the information, and dig up all the holes regardless. This is because Normal will have to incur this information cost one hundred times. Yet there is only a one percent chance that it will cause him not to dig up a hole. If Normal incurs the information cost, then, in retrospect, this was optimal with regards to any hole that did not contain a coin. Yet, this cost was incurred at all holes that did contain a coin, without any corresponding saving to counteract this rising cost. Hence, in judging the price of the information, all instances where it is gathered need to be considered, not just the instances where it changed behaviour.

---

[94] In the language of economics, this could have been expressed by saying that computational power is a finite resource that needs to be distributed over infinite needs ("options"). For Normal the opportunity cost of code-breaking rose, hence he redistributed this resource.

This reinforces the main conclusion reached above with regards to information cost in general. It might well be optimal to follow a rule that will occasionally result in unsuccessful action (i.e. digging unnecessarily) than to simply act so as to prevent unsuccessful action.

4. Conditions under which Rule-Following is Optimal

The main advantage of rule-following lies in the fact that it saves computational power, which can then be used elsewhere. It does this by telling the actor what criteria to act on. In doing so it tells the actor what to ignore. It was argued that there are three factors that serve as a rational base for rule-following. The first concerns the case where the information that can change my action is unobtainable, i.e. ignorance. The second is if the information that can change my action is too expensive, i.e. information cost. The third, which is really a special case of the second, is if the information that can change my action comes at too high a computational cost. Hence a utility-maximiser should follow a rule under the following conditions:

1. A choice recurs that is alike in a certain respect. Example: Normal has to decide whether to dig or not dig.
2. Consistently choosing one option is demonstrably superior to consistently choosing the opposite. Example: Normal is better off always digging than never digging.
3. The instances where the option that is not consistently superior is actually superior either cannot be known in advance, or are too expensive to determine. Example: Normal was portrayed as unable to figure out the workings of the pseudo-random number calculator, or this would have been too much effort, etc.

In the following chapter I will attempt to show that certain choices in ordinary life satisfy the above three conditions. Hence, with regards to these choices, a utility-maximiser should follow certain rules.

But first the account of rule-following needs to be extended to the fourth possible source of rule-following. This source is, in many ways, the most fundamental. It also

delineates a realm of action where rule-following is superior to calculation, but does not stop there. It will be argued that this factor does not only demonstrate the optimality of rule-following *vis-à-vis* calculation, but actually shows the impossibility of being anything other than a rule-following being. This argument can be made by arguing for a specific understanding of the relation between rational choice theory and bounded rationality.

5. Rational Choice and Bounded Rationality

The essence of rational choice theory is that it views actors as maximising utility. This means that actors consider the different options available to them, and then choose the one that maximises utility (Monroe, 2001: 153).

In a series of classic papers starting in 1948, Simon developed his conception of "bounded rationality" as a critique of, and corrective to, the theory of rational choice. Bounded rationality emphasises the constraints on human decision making[95] in order to show that people develop certain strategies for dealing with these constraints. One such a strategy is what Simon called "satisficing".

> A decision maker who chooses the best available alternative according to some criterion is said to optimise; one who chooses an alternative that meets or exceeds specified criteria, but that is not guaranteed to be either unique or in any sense the best, is said to satisfice (Simon, 1997: 295).

The dispute amounts to the following: According to Simon an actor will often act once she finds an alternative that meets a given criterion, according to rational choice an actor will act after choosing the best option.

It is here that the prior discussion regarding calculation and rule-following is of relevance. A "satisficing criterion" is a criterion that tells an actor to act once she finds an option that has a certain characteristic. Hence the above discussion of rule-

[95] Simon notes that economics has never been oblivious of the ideas behind bounded rationality. Both Lucas and Keynes can only introduce business cycles into their theories by assuming that the labour

following was also a discussion of satisficing. A rational choice-maximiser is an actor who chooses the best option relative to a choice-criterion. Hence the above discussion of "calculation" was also a discussion of rational choice "maximising". Simply put, rational choice is concerned with what was defined as calculation, bounded rationality with what was defined as rule-following[96].

What is the relation between rational choice and bounded rationality? The view that is most often taken is that bounded rationality can be seen as an instance, not a refutation, of rational choice. In other words bounded rationality can be seen as rational choice under informational and computational constraints[97]. It is easy to see how this conclusion can be reached. Consider Normal's choice to simply dig up all holes in the above example concerning computational cost.

Normal was following a rule ("satisficing"), but this rule was not followed without a reason. He calculated the difference between following a rule and calculating, and decided to follow the rule. Or, simply put, he calculated and discovered that he should stop calculating. But his rule-following did not represent the final word with regards to an explanation of his action. Rather the rule was the *result* of calculation, and it is only when his *initial calculation* is taken into account that his actions are fully explained. In this way it can be said that his rule-following was the result of calculation, which is the primary explanation of his action. Or, in other words, satisficing is an instance of maximising, and explicable in terms of it.

Below I will argue that, while the above analysis is not without merit, there is an important sense in which it is wrong and/or incomplete. I will argue that, when it comes to the relation between rational choice and bounded rationality, the above analysis has things exactly backwards. Instances of bounded rationality cannot be

---

force (Keynes) and businessmen (Lucas) have limited knowledge and decision-making capacities, causing both to suffer a variant of the "money illusion" (Simon, 1997: 294).

[96] This is not one of the standard definitions of rule-following and calculation. Nothing substantive depends on such stipulations. The definitions of rule-following and calculation were specifically chosen so as to correspond to the fundamental difference between rational choice and satisficing.

[97] Simon seems somewhat ambivalent about this interpretation of his work (1997: 296). See Elster (1986: 25-27) for a discussion of, and argument against, this interpretation. In contemporary writings of bounded rationality theorists this interpretation is often explicitly rejected. Gigerenzer and Selten, for example, explicitly call the view that bounded rationality is optimisation under constraints "inappropriate and misleading" (2002: 5).

explained with reference to an earlier case of rational choice. Rather all instances of rational choice are to explained with reference to earlier instances of bounded rationality. It will also be argued that the above reversal is more than a case of "six of one, half-dozen of the other". In other words it will be claimed that there is a sense in which rules are more fundamental than calculations.

One way that the argument can be made is with reference to the difference between the way human beings play chess, and the optimal method of playing chess.

6. Rules and Calculations: Kasparov versus Deepest Blue

6.1 The Game

Consider the differences between the way in which a human expert (call him "Kasparov") plays chess, and the way in which the greatest conceivable computer (call it "Deepest Blue") would play chess. The amount of possible moves in chess is, at any given state of the game, finite. The amount of moves resulting from any possible move is similarly finite, etc. This means that the amount of possible games is, while incredibly huge, finite. Imagine that Deepest Blue has a computational capacity that allows it to run through all possible moves within the given time constraint, and that this is the method it uses in playing chess. Deepest Blue is, simply put, the ultimate lightning calculator.

Kasparov cannot possibly play chess solely by using lightning calculation. The human brain, while an impressive information-processing device[98], does not have sufficient computational capacity to make this possible. Rather he would have to use his "expertise". Simon describes the "expertise" of brilliant chess players as consisting of "selective heuristics" (1997: 185). In terms of the above analysis this amounts to following a complex set of rules. The difference between the way Deepest Blue plays chess and Kasparov plays chess allows for a good illustration of the relation between calculation and rule-following, and hence rational choice and bounded rationality.

---

[98] The average processing rate of the brain, according to Dennet (1987: 328), is orders of magnitude faster than the fastest supercomputer. Even this is not sufficient to play chess by calculations, as the hypothetical Deepest Blue does.

If the "intentional stance" is taken towards Deepest Blue it can be treated as an agent with the following characteristics. All choices made by Deepest Blue are attempts to maximise utility, where utility is something like "chance of winning". It does this by considering all possible options, and chooses the one which best maximises utility.

Assume that Deepest Blue does not make the stupid type of computational mistake that we all fall victim to from time to time. In other words, it functions perfectly, and knows this with absolute certainty. Hence it never needs to go back and verify the result of its calculations.

Deepest Blue gives content to the idea of a perfectly rational actor. He is, as the term was defined earlier, an absolute calculator. In this sense he is similar to the hypothetical Perfect that was used to illustrate some points above. Deepest Blue does not have instantaneous computational capacity like Perfect, yet this does not matter. Deepest Blue has, at bottom, only one goal that decides all his actions – to win at chess. Time is not a factor, Deepest Blue is indifferent between winning within an hour and winning in two months' time. The possible courses of action open to Deepest Blue are the possible moves in chess, and hence are finite. Perfect was assumed to view time itself as a valuable commodity, and to have infinite options. This necessitated instantaneous computational capacity in order for Perfect to be an absolute calculator. Dropping these two requirements means that Deepest Blue can be a perfect calculator without having instantaneous computational capacity.

Deepest Blue is also operating in an environment where, with reference to the factors he cares about, changes are interspersed with periods during which his environment stays constant. Simply put, the pieces are moved and then remain static until they are moved again. This was not the case with Perfect. It was assumed that she is operating in a constantly changing environment.

Hence Deepest Blue is the ultimate rational actor, or absolute calculator. How will Kasparov's method of playing differ from that of the perfectly rational actor?

6.2 Kasparov's Method

The central way in which Kasparov differs from Deepest Blue is that he does not have the computational capacity to run through all possible games in the time allowed for a move. There is no reasonable time-constraint that will make this a possible option for Kasparov, unless we stipulate that he is not immortal. Rather he will have to explore the consequences of the moves that, based on his considerable expertise, appear most promising. Presumably Kasparov will look for options that have certain properties, and if he finds these properties, explore the possibilities of utilising the option[99]. The properties or criteria whereby the notion "most promising options" is defined will be one or more satisficing-principles, or, in other words, rules. Hence Kasparov will satisfice in choosing which options to explore.

The exploration of these options will, however, take time. Time spent exploring a bad option is time stolen from exploring a better option. Kasparov will again, based on his expertise, "know what to look for" in a potential move. In other words there is a set of properties that guide him in deciding whether to continue or stop the exploration of a given option. Again this "set of properties" can be viewed as a satisficing principle, or rule.

Kasparov will follow rules, or satisfice, in choosing *which* options to explore. He will also follow rules ("search-rules" as Simon calls them) in choosing whether to continue or discontinue his search. It is also possible, within a game, that Kasparov can doubt the rules that he is using, and try and determine whether any given rule or criteria might not be misleading. For the vast majority of the complex set of rules he is following, however, this cannot be the case. Rather these rules are something he brings with him to the game, rather than something he decides to use during the game.

Hence there are at least three important senses in which Kasparov will follow rules. He will not explore all options, but pre-select from them. He will use search-rules. And these rules themselves will simply be something he brings with him to the game,

---

[99] In a study that asked chess grandmasters to describe their thought-processes while solving chess-problems it was found that they considered eight likely moves, and progressively "played out" five of these to see what possibilities they opened up (Klein, 2002: 115).

not issues decided during the game. It will only be within the context of rule-following that he will calculate. In other words, once he has committed to the most promising moves, and once he has decided how far in advance to plan, and once he has decided how to judge the merit of the promising moves, only then will he calculate *vis-à-vis* these options and criteria.

This illustrates an important principle. Calculation is only possible once the options, choice-criteria, and evidential-criteria whereby it is judged which option is most likely to fulfil the choice-criteria, have been fixed. These are all matters that, with reference to a specific case of calculation, factor in as something the actor is already committed to, i.e. a rule.

6.3 Relevance of the Difference Between Kasparov and Deepest Blue

A defender of rational choice theory might well wish to defend it in the following manner. It was already observed that a rational actor who can incur computational costs would sometimes decide to follow a rule so as to avoid computational costs. This presents the possibility that all the rules followed by Kasparov can be interpreted in this manner. In other words all his instances of rule-following can be seen as flowing from cases of calculation. In this manner bounded rationality becomes a special case of rational choice.

This will not work, for at least two (related) reasons. Any prior calculation that resulted in the decision to follow a rule will again be based on other rules. This does not result in a "chicken-or-egg"-situation. The possibility of starting with a "pure" calculation devoid of rules is unintelligible for any being with finite computational capacity.

Imagine a being had to, based on an effort to avoid computational costs, try to calculate which rules to follow. In doing so it is trying to calculate whether to calculate. But this calculation is one that he himself might be better off not making. This means that the rational actor would now have to calculate whether to calculate

whether to calculate[100]. Here an infinite regress arises, for each of the actions of calculation need to be justified as flowing from a prior decision to calculate. This is logically impossible. Rather the actor would start with a commitment to some calculation, and work from there. This means that whatever is taken to be the "first calculation" would have to start from within a context of rules that are already followed[101].

The second, and not unrelated, problem with the idea of reducing all rules to maximising calculations concerns the issue of time[102].

Imagine a being that is not, because of the self-referential problem discussed above, excluded from being an absolute calculator. But her calculations do take time. While she is trying to make her "first and absolute calculation" the clock will be ticking away. If this being is sufficiently like a human being, so that the options confronting her are *infinite*, she will never get around to the first decision.

It can even be granted that the above problem is, somehow, overcome. Imagine that the above being has somehow found a legitimate means to reduce her available options to a finite set. Now she has to calculate the utility of the remaining possibilities. Even grant that during this process she can somehow hit upon search-rules and evidential criteria that can be related back to pure calculations. Meanwhile the clock is again ticking away, with opportunities fading as new ones arise. In refusing to "blindly" enter a state she is already in a state – a state of dallying. This is surely not an optimal way of acting. If she had simply seized some option that seemed okay based on one criterion, e.g. "satisficed", utility would probably have been

---

[100] Or, as summarised in Gigerenzer and Selten (2002: 5): "[T]he cost-benefit computations are themselves costly, and demand a meta-level cost-benefit computation, and so on".

[101] The above problem arises with reference to deciding *whether* to decide. It also seems to arise for the related problem of *how* to decide. The question as to *how* to decide has been recognised by some authors as resulting in regress problem. In an article that attempts to overcome this problem for a special case, Smith (1994) mentions Raiffa, Rawls, Elster and Resnik as among the authors who feel that the infinite regress also arises when deciding *how* to decide. Russell Hardin is specifically mentioned as attempting to overcome this problem by starting with the idea of satisficing, as opposed to maximising (1994: 196).

[102] These problems do not arise for Deepest Blue, despite the fact that his computational capacity is not instantaneous. This was due to a number of factors peculiar to himself, the most important of which is that time is not a commodity or price to him. The other factors were explained above. The case of Deepest Blue is sufficiently unlike that confronting a human being with regards to all these factors that I am excluding it from further consideration.

optimised much more efficiently. If it is agreed that economists are correct when asserting that five-dollar bills are rarely left lying on the side-walk, i.e. that obvious opportunities are rarely missed, it must also be agreed that she would have satisficed.

The point that the above seeks to illustrate is that a being with finite computational capacity must be a rule-follower before it can be a calculator. The idea of calculation "all the way down" falls apart on grounds of coherence, when the influence of time is considered, etc. Even if the idea of starting out as calculator can, in some sense, be saved, it can be a decidedly sub-optimal way to act. The reasons explained above all point to the implication that all instances of rule-following cannot possibly be seen as the result of prior calculation.

6.4 Can Rule-Following Become Calculation?

The above argument against a being with finite capacities *starting* as rational actor does not necessarily result in the conclusion that such a being cannot *become* a rational actor. Rather a somewhat intriguing possibility presents itself. What about a being that starts out as a rule-follower, but becomes a calculator? Is it possible for satisficing to become a rational choice?

Consider the following. "Boring" is a being with only one goal. This goal is to win at tic-tac-toe. For whatever reason Boring is not allowed to figure out the rules of the game prior to playing his first match. Rather he is just thrown into playing tic-tac-toe, and has to figure out the rules as he goes along. In other words, Boring starts as a rule-follower.

Tic-tac-toe is a rather simple game. Once one has figured out a few simple "strategies", it loses its appeal because one can no longer lose. Assume that Boring figures out these strategies. Also assume that Boring's opponent makes moves that are random, so that Boring cannot, even in principle, outwit it. Boring can only make the best move allowed by the game. It will win some games and draw the rest.

All the information needed for any given choice is either freely obtainable by Boring, or impossible to gain because of randomness. Hence he will simply choose the best

option consistently. In doing so Boring will move from being a satisficer to being a rational actor.

It does seem that, despite starting out as a rule-follower, it is possible to become a rational actor. But, I think it reasonable to suppose that this is not the case with *homo sapiens*. Just consider the number of ways that we differ from Boring. For a start, life is, at any moment, infinitely more complicated than tic-tac-toe. Matters are made even more difficult by the fact that the world, in contrast to tic-tac-toe, is constantly changing. Furthermore, our "utility" can be affected by an inexhaustible number of factors, i.e. we care about an inexhaustible number of things. Boring was assumed to only care about winning at tic-tac-toe. Simply put, life simply does not become that much easier that quickly, in the sense that tic-tac-toe does.

The same point can be made with reference to Kasparov. Kasparov, despite his intelligence and concentration, will never reach the point where he can play chess by calculation. There is simply a difference of orders of magnitude between the amount of computational power possessed by Deepest Blue and the amount possessed by him. It is probably fair to say that the average person, in ordinary life, is even more orders of magnitude below Kasparov-playing-chess than Kasparov is below Deepest Blue.

The amount of concentration and intelligence that Kasparov brings to chess probably represents the zenith of possible human computational ability, or something close to it. So any satisficing that Kasparov needs to do will be magnified tremendously for the average person confronting life. Kasparov has to choose from a set of possibilities that are, in principle, finite. The average person has, at any given point in time, an infinite number of possible actions. In choosing between these he does not have nearly the intelligence or concentration that Kasparov exhibits in playing chess. In fact, neither will Kasparov. It is probably correct to guess that, while playing chess, he devotes nearly none of his intellectual resources to any problem other than winning. Kasparov is further helped by the fact that chess is a game with defined rules, where the pieces are only moved occasionally. Metaphorically speaking life is probably best described as a game where the rules can change at any moment, and the pieces never hold still. Hence the idea of learning what rules of action are optimal, and then

retroactively judging them, as done by Boring, has limited application to human action.

Rational choice theory attempts to start with calculations, and then accommodate rules as a special case of a prior calculation. This can be a useful way of looking at action, as will be discussed below. But, as a description of human action, this seems to be the wrong way around. Rather the idea of rule-following seems to be the most fundamental category of human action. This is however, not an *a priori* truth about the nature of a possible actor. Rather it is the direct result of human beings having finite computational capacity, infinite options and having to make all our decisions while the clock is ticking.

Until now, nothing much in this chapter has depended on any particular theory regarding how the mind works. All the conclusions reached above are supposed to follow logically from the assertion, surely uncontroversial, that human computational power is a finite resource that needs to be distributed over potentially infinite needs. It was argued that any being that needs to do this cannot possibly be a rational actor, because of the infinite regress problem, etc.

But there is another factor besides cognitive constraints as such that speaks against the idea that man can be a rational actor. This will become apparent if it is considered how we came to have the minds that we do.

7. Evolutionary Psychology and the Modular Mind

7.1. The Evolved Mind

It has already been stated that the mind is an information-processing mechanism that allows a utility-maximiser to attempt to maximise utility. Logically it follows that, if computational capacity is unconstrained, the best mind to have would be a generalised information-processor that is constantly following the following decision-rule: "Determine which action will maximise utility and then choose it". The best logical

possibility would be a mind with infinite computational capacity, always following the above decision-rule.

However, the constraints of the physical world make this impossible. Computations take time; this fact by itself has some interesting implications that were discussed earlier. Rational choice tries to get around this problem by picturing the subject as a constrained maximiser. In other words, the mind is, fundamentally, only following rules because they are the result of earlier calculations that determined these rules to be optimal. It was already argued above that there are some fundamental problems with this view. The greatest of these problems is that this idea is logically incoherent because of the infinite regress problem. There is, however, another objection to this idea that I consider to be equally devastating. This is that the type of mind needed for rational choice theory to be an accurate description of human action cannot possibly have evolved.

The next part of this chapter rests on the empirical claim that the mind is the result of evolution. This claim is taken to be common knowledge in evolutionary psychology, and will not be defended here[103]. The claim that will be defended, however, is that an evolved mind is extremely unlikely to belong to a rational actor. This can be explained by considering the type of actor that evolution can be said to be.

7.2. Evolution as Bounded Rationality Decision-maker

Evolution can metaphorically be viewed as a designer that attempts to maximise utility, where utility is defined as "differential reproductive fitness". However, evolution as a designer can only work according to a very simple decision-rule. It follows the satisficing criterion that any change, arising from mutation, that increases relative reproductive fitness, is selected.

[103] For an excellent introduction to evolutionary psychology, see Badcock (2000). For a consideration of the evidence for the "modular mind", see, for example, Gigerenzer and Selten (2002: 83-102).

This constrains decision-making in several ways. Firstly the potential options that can be decided between are pre-selected. A mutation, no matter how beneficial, can only be selected after arising through the chance process of mutation. Hence a situation arises that is similar to Kasparov's pre-selection of alternatives, explained above.

There are also some options that cannot be utilised by evolution. If a potential mutation would decrease fitness, but, in conjunction with a subsequent mutation enhance fitness a great deal, then such an "enabling" mutation cannot be selected. Design-processes that consist of more than one step can only be selected if each individual step is beneficial.

In other words, evolution is an extremely myopic actor that can only see one "move" ahead - and then only those "moves" determined by a random process. This view of evolution as a boundedly rational actor explains the seemingly perverse nature of much biological design. Where efficient human design is normally characterised by simplicity and elegance, evolutionary design tends to be intricate and complex, yet effective. This has lead several authors to compare evolution to a "backwoods mechanic" that makes ingenious use of seriously limited materials through sheer cunning rather than elegant design principles.

Consider, for example, the eye. On the one hand it is an ingenious device, our "...experience of the colours of objects depends on a process of visual analysis that, though largely unconscious, must be highly sophisticated and complex" (Shepard, 1992: 495). On the other hand, it seems to be built the wrong way round. The "blood vessels and nerves are on top of the light receptors and this not only obscures the view, but creates a tendency for the retina to become detached... An altogether better design would be the obvious one of having the wiring and plumbing behind the retina, which is what is found in molluscs like the squid and octopus" (Badcock, 2000: 19). This peculiar design exists for no better reason, probably, than the historical fact that "the light-sensitive skin cells from which the vertebrate retina originally evolved were under the surface of the skin, rather than on top of it" (2000: 19).

This seemingly incongruous mix of genius and perversity is often the distinguishing mark of evolutionary design as such. And, if it is granted that the mind is the product

of evolution, then it should be characteristic of the information-processing rules implemented in the brain as well.

7.3. The Modular Mind

Taking their cue from the nature of evolutionary design, evolutionary psychologists have proposed a view of the mind that sharply contrasts with the unified, all-purpose general calculator that is often assumed in writings about the mind. This is replaced by the "Swiss army knife" model, also known as the "adaptive toolbox" or, simply, the modular mind. The following is a fairly standard description of the main tenets behind the idea of the "modular mind".

> The picture that is emerging from both noninvase studies of normal brain function and from clinically defined syndromes resulting from brain damage from strokes, injury, and neurodevelopmental disorders is one of different neurological substrates serving different cognitive functions. This picture has provided philosophers a glimpse into the possibility that Descartes was mistaken about the unity of consciousness. The neurological divisibility of mind also provides the key to understanding its evolution. *The Cartesian view of a seamless whole makes it hard to see how such a whole could have come into being, except perhaps by an act of divine creation.* By recognising the modularity of mind, however, it is possible to see how human mentality might be explained by the gradual accretion of numerous special function pieces of mind (Cummins and Allen, 1998: 3, my italics).

The idea behind the modular mind, is, simply, that evolutionary design would be incapable of providing us with an all-purpose general calculator. Rather genetic mutation influences the design of the brain, and this effectively creates different decision-rules. These decision-rules are then selected by evolution based on their effect on the reproductive fitness of the individual.

These decision-mechanisms tend to be an "adaptive toolbox", i.e. context-specific information-processing rules that get the job done. These mechanisms

relate to a specific function that can be seen as sub-goals of an effective maximiser of reproductive fitness. For example the landmark collection *The Adapted Mind* (1992), edited by Barkow, Tooby and Cosmides, discusses different "mental modules" related to, among others, social exchange, sexual attraction, language and our aesthetic sense that can be seen as specific adaptations in response to certain evolutionary pressures.

This "adaptive toolbox" view of the mind has fascinating implications that fall outside the main ambit of this thesis. The main point relevant here, and one that follows from the assertion that the mind was designed by evolution, i.e. a boundedly rational actor, is that we have no reason to expect biology to agree with logic. In other words, the optimal solution to a specific problem, given by a mathematician or logician, will only agree with the satisficing solution, provided by evolution, by sheer fluke[104]. Hence, for the most part the innate information-processing mechanisms provided by evolution will be rules, and not calculations.

Two reasons why rational choice theory cannot be the final truth concerning human action have now been discussed. The first is due to the fact that computational capacity is a finite resource that needs to be effectively distributed over infinite needs. This implies the existence of certain trade-offs between accuracy and computational cost[105] that gives rise to a sense of rule-following irreducible to rational choice. The second is that the mind is partly composed of innate decision-mechanisms, and that these have been programmed by evolution. Evolutionary design is constrained in such a manner that optimal results are highly unlikely.

---

[104] The same point is made by Dennet (1987: 51): "But not only does evolution not guarantee that we will always do what is rational; it guarantees that we won't. If we are designed by evolution, then we are almost certainly nothing more than a bag of tricks, patched together by a *satisficing* Nature... The demands of nature and the demands of a logic course are not the same.... [T]here has probably been some positive evolutionary pressure in favour of "irrational" methods".

[105] How much accuracy can very simple rules provide? In Gigerenzer et al. (1999) one-criteria decision-making and other simple strategies are matched against complex decision-making strategies like multiple regression. The counter-intuitive conclusion reached is that vast savings in computational cost can come at very little loss in accuracy (and sometimes even a gain where "noisy" data is encountered and phenomena like overfitting occurs). This is due to "ecological rationality", simple strategies exploit the structure of information in a given environment to do rather well. For instance,

These two reasons are independent, but mutually supporting reasons why rational choice theory is mistaken. It now remains to briefly relate these ideas to the hedonism being defended in this thesis.

8. The Learned, the Innate, and the Evolution of Preferences

It is surely uncontroversial to assert that some of our information-processing mechanisms are learned and that others are innate. The innate mechanisms were argued to be the result of a boundedly rational actor (evolution) trying to maximise reproductive fitness. The learned mechanisms were argued to be the result of a boundedly rational actor (*homo sapiens*) trying to maximise utility.

Innate information-processing rules can either determine a process completely (be "hard-wired" into the brain) or increase the likelihood of it occurring by utilising some proximate mechanism. Most innate decision-making rules surely belong to the first category in that they operate automatically, but this is not the case with those rules most directly linked to human action.

An evolutionary explanation will normally explain the adaptive significance of a certain mutation. This does not, however, give the full story when it comes to evolutionary explanation. An explanation also needs to be given of the proximate mechanism involved, i.e. of the way in which the evolutionary goal is accomplished. By what proximate mechanism can evolution get a utility-maximiser to do its bidding? The answer that suggests itself is the manipulation of preferences. Those preferences that maximise reproductive fitness can be expected to evolve at the expense of those that do not. In previous chapters it was argued that preferences are ultimately to be explained with reference to pleasure-maximisation. If this is the case, then evolution controls behaviour by letting those beings evolve that gain pleasure from activities that happen to maximise reproductive fitness[106].

---

simple one-criteria decision strategies will do well in environments where data is "skew", i.e. has an L-shaped distribution (1999: 124).

This implies that there are two distinct ways of potentially explaining the actions of a utility-maximiser. One would be an examination of the actions that would, given cognitive constraints, best satisfy his preferences. The other would be with reference to an account of the evolution of these preferences. Hence, while economists treat preferences as a given, evolutionary psychology can go beyond this. This is an opportunity that I will exploit in chapter 6 when the topic of "altruism" is discussed.

In this chapter it was argued that man, necessarily, is a rule-following creature in a sense much more basic that the one in which he is a calculator. In the following chapter I will try to formulate one such rule that should be followed by any utility-maximiser with our cognitive constraints.

One question still remains to be discussed. If man is a boundedly rational actor, does this imply that rational choice theory is useless? Below I will give an argument for why I believe that rational choice theory still has a place in the explanation of human action.

9. The Status of Rational Choice-Explanation

9.1 Applicability of Rational Choice

Rational choice theory can still, despite its shortcomings, serve as a useful device to explain and predict human action. This can be illustrated in the following manner.

Imagine a group of beings with the intellectual capacities that are expected of intelligent people. They know the rules of chess, but are complete novices at the game. Now imagine that they play a round-robin tournament against each other in a game with extreme time-limits, where only 10 seconds is allowed per move. Further imagine that there is practically no time between games, and that they do not get tired. In other words these novices are playing what amounts to continual chess. What can be expected of their quality of play?

---

[106] The idea that evolution influences behaviour through pleasure/pain as proximate mechanism can be found in Darwin, Spencer, William James and others. For a discussion see Badcock (2000: 125-129).

The novices would start by following extremely simple strategies, and with widely differing results. Those who happen to hit on a good strategy early on will be very successful, those who do not will lose badly. It is unlikely that any of the strategies will be, at the start of the tournament, very good.

Over time, however, the logic of any evolutionary process will take hold. The bad players will learn from the good players so that the quality of strategies will become less stratified. The quality of the winning strategies will increase as time elapses. If this goes on long enough the games will reach a point where the chess that is being played is, considering the time constraint, rather good. It will probably progress to a point *well beyond the strategy of a being with much greater computational power* who was forced to start playing in a similar manner, but who has not had the advantage of having his strategy improved by evolutionary trial-and-error.

If it is known that the above had occurred, then there are two methods of determining the strategies that will eventually be followed. One method can be realistic, and take the capacity of the players, the amount of games, etc. into account. The drawback to such a method is that it needs an awful lot of information regarding such detail in order to produce useful results, and will be incredibly complicated.

The other option would be to simplify the question in the following manner. It can simply be asked: If a utility maximiser has to play one game of ten-second chess, what strategy will produce the best results? This question attempts to relate a set of rules (the strategy) back to an original maximising decision. There is every chance that the answer to this question will be very close to the strategies produced by the evolutionary process. It does not matter that the ability to come up with such a strategy as a matter of rational decision is beyond the cognitive limits of the actors in the round-robin tournament. *Procedurally* it would not be an accurate description of the process that led to these strategies, but the *outcome* would be similar[107].

[107] This distinction between "procedural rationality" and "rationality of outcome" is due to Simon. Procedural rationality refers to the process by which decisions are actually made. Rationality of outcome (or, "substantive rationality" as Simon refers to it) refers to the outcome that best maximises utility. (See Simon 1997: 25-26).

This leads to the conclusion that rational choice theory has two main uses. It can, firstly, be used as a descriptive tool where an actor was performing relative calculation. In other words it can be used where it is reasonably certain that an actor had pre-selected certain options, criteria, etc. so that the decision was, within this context, the result of calculation. Here it applies, both as an account of procedure and to predict the outcome of a decision. Secondly, it can be used to predict the outcome of a process that, due to its logic, will tend to converge on the most rational outcome. Here it is procedurally incorrect, but can be a useful predictive tool. It has the advantage that it need not take cognisance of the complexities regarding actual human decision-making.

It should be possible to give an account of the different factors that determine when, and to what degree, the above method will be accurate. This would include factors like time, rate of learning of the actors involved, etc. This chapter will not attempt to present such an analysis. What is important to note, however, is that the applicability of rational choice is not an *a priori* matter for analysing all decisions. Rather it would depend on the existence, in a specific situation, of the qualities that would be the subject of the analysis referred to above. Whether it would get the answer right would also depend on the usual intricacies of evolutionary processes of the type mentioned earlier, i.e. path dependence[108], the possibility of falling into local maxima[109], etc.

9.2 Example of Applicability of Rational Choice - Hobbesian Contractarianism

Below I will discuss an example of how rational choice can be used in the explanation of human action. This will be done with reference to some of the objections traditionally made against Hobbesian contractarianism[110].

---

[108] A process is path dependent if the possible outcomes are highly sensitive to the initial steps in the process.

[109] Sometimes it is necessary to take two steps backward in order to take five steps forward. A process that cannot "backtrack" in this way can get stuck in a "local maximum" – a state that is sub-optimal, yet optimal compared to all states reachable in one move. This is precisely the problem with evolutionary biological design explained earlier.

[110] I wish to insert the disclaimer that the choice of Hobbes as an example has nothing to do with the main aims of this thesis. While Hobbes's egoism fits in well with what is being argued for here, I do not wish to defend contractarianism in this thesis as such. Hobbes is chosen here only because I think that some of the criticism against him misses the point in a way that illuminates the relation between rational choice and bounded rationality.

Hobbes views society as the rational creation of self-interested individuals who wish to avoid the state of nature. These individuals cede their freedom to defend against aggression to a sovereign who then exercises this power on behalf of the individuals who ceded it. This agreement then constitutes the Hobbesian "social contract", which is presented as the notion upon which society is founded.

One objection to Hobbes that is often encountered is that the idea of a "contract" cannot be an accurate account of how a society actually comes into existence.

> There never was such a contract because there never was a time when men lived [without] a society of some sort (Campbell, 1981: 85).

> Language ... is a rule-based activity requiring rule-based learning and authoritative standards, and yet men cannot make contracts without using language; therefore it must be nonsense to think of contracts being made in the state of nature (Campbell, 1981: 86).

The above objections concern both the actual history of contracts and the logical coherence of presenting it as a founding notion. "Contract" didn't serve as the basis of society *historically*, nor can it *logically* serve as the basis of society. These two objections correspond to the objections made in this chapter to seeing bounded rationality as a special case of rational choice. Historically it is not the case that there ever was a time when man was not already acting, i.e. already a rule-following creature. In similar way it is not the case that there ever was, according to the above objection, a Hobbesian "state of nature". Logically the idea of founding all rules upon previous decisions is incoherent because of the self-referential nature of computational costs and the infinity of possible options. In the same sense the idea of founding a society on the notion of a "contract" is incoherent, since a contract already presupposes social institutions like language.

I would contend that Hobbes can be partially defended against these objections in the same way that the use of rational choice theory was defended above. In fact, the above considerations must make it reasonably clear that Hobbes is giving a type of "rational choice" explanation of society. The above objections point out that

"contract" cannot have served as the historical basis of society. Hence the analysis Hobbes gives is, in terms of the earlier discussion, *procedurally* inaccurate. But the portrayal of procedural rationality is not the only possible use of the theory of rational choice. It can be a useful shortcut for identifying the most rational outcome of a process. And, provided the actual process whereby a society comes into existence has the type of qualities that allow for the evolution of the nature of society, the answer of the rational choice method might well be close to the outcome of the actual process.

There is, as pointed out earlier, good reason for using such a method. The actual process of how a society comes into being is no doubt filled with very complex processes, depends on historical contingencies, and is inordinately complicated. No doubt there are a lot of the historical facts regarding this process that we will simply never know. The procedurally accurate way of portraying this process would probably have to include something like a large number of rule-following beings that interact in situations that can be modelled as games with certain equilibria, with external factors often changing the possible outcomes, or the game itself. This type of analysis would be extremely difficult to give in full[111].

Here rational choice proves its worth by determining the most rational outcome of the process. Such a reconstruction can have valuable predictive and explanatory power. Predictive power in that it can give a good approximation of the eventual outcome. Explanatory, in that, if a certain outcome is observed, and if this outcome is also the one predicted by rational choice, it is reasonable to suppose that the actual process operated under the constraints of evolutionary logic.

Hence Hobbes can be defended if his claims are understood as a reconstruction, according to rational choice theory, of a process that lead to an outcome that matches the one predicted by rational choice. This means that his analysis, even if it is conceded to be *procedurally* incorrect or incoherent, is far from being without value.

---

[111] The closest thing to such an analysis that we have is probably the magisterial, two volume *Game Theory and the Social Contract*, by Binmore (1994; 1998). Yet this is, for obvious reasons regarding the necessary gaps of our knowledge in areas like history, psychology, etc., still a massive simplification of the actual process.

9.3 Relevance of the Debate between Rational Choice and Bounded Rationality

Above it was argued that an understanding of the issues involved in adjudicating between rational choice and bounded rationality allows for a deeper understanding of some of the traditional objections to Hobbesian contractarianism. This should come as no great surprise, whenever one deals with human decision-making or action in any way the topic will probably be implied somewhere. This issue is often relevant in the most surprising ways. I will give one more example of this debate's relevance, and from a field that does not seem to have much use for the writings of economists or biologists.

It was argued above that realising time to be a finite resource has certain direct implications for our view of human action. It is probably no accident that there are certain parallels between viewing man as a habitual satisficer and the view of man put forward by the philosopher most famous for his writings regarding the relation between human existence and time – Heidegger. Below I will briefly touch upon some similarities between Heidegger's criticism of traditional notions of subjectivity, and the bounded rationality criticism of rational choice theory.

Heidegger rebels against the idea of man as a being who can make decisions, and only then engage in the world[112]. Rather he speaks of "man" as *Dasein*, literally "there-being", in order to emphasise that man is always already an inhabitant of the world. As such man does not encounter entities, which subsequently acquire meaning. Rather "the world presents its significance to Dasein" (Gelven, 1970: 53). In terms of the analysis contained in this chapter this seems analogous to saying that I am always already interpreting my world according to a set of rules, i.e. it already has a meaning to me. I am not encountering the world, and then deciding how to interpret it.

It is within the above context that Heidegger distinguishes between the "ready-at-hand" and the "present-at-hand". If something is "ready-to-hand" it appears in terms of its "equipmentality", i.e. the functional relationships it has relative to a particular situation. If something is "present-to-hand" it appears in its "thinghood". An example

[112] "The question of existence never gets straightened out except through existence itself" (Heidegger, 1987:33).

of Gelven's (1970: 199-200) can clarify this distinction. A heart-surgeon trying to save a life sees a heart in terms of its capacity to sustain life. This is not necessarily the case with a theoretical biologist. She will see the heart as a "thing with properties", rather than as a set of functional relations[113].

In our everyday encountering of objects they are generally viewed as "ready-at-hand". This only changes if an object suddenly does not satisfy the conditions that allow me to deal with it in such a pre-reflexive manner. In other words, if my key breaks in the lock I will suddenly cease to see it in terms of "opening a door", and start seeing it in its "thinghood" in order to find a way to view it that will work in this new situation.

This type of view seems consistent with the view of rule-based activity developed in this chapter. We are always already interpreting the world, different non-contradictory ways of looking at things are weighed in terms of the goals of the actor. These interpretations are useful in that they suggest a way to deal with objects, i.e. "a key is something that opens a door". In this way our perception of the world already suggests certain possibilities and make others (like: "the key can help steady an uneven table") less likely to occur to us. These habitual ways of interpreting will tend to be upset only once something goes radically and obviously wrong. Here the object is consciously considered in terms of a wider array of qualities, until a new way of looking is found that again allows for it to be dealt with in a manner that requires little or no conscious attention. To summarise: people are always already following interpretive rules in dealing with their environment ("ready-at-hand"). They only introduce a degree of contextual calculation ("present-at-hand") in order to find a new rule when the previous rule is obviously failing.

I do not wish to make any ambitious claims regarding the relationship between the claims of this thesis and Heideggerian philosophy. What I wish to claim is modest, but not uninteresting. The rational choice view of human agency, i.e. the view that human action springs from consciously calculating the best option and then acting

[113] It might be objected to Gelven's example that a theoretical biologist is even more likely to see the heart "functionally" than a surgeon trying to save a life. Functionalism is, after all, a dominant tradition in biology. Whether the specific example that Gelven cites is correct does not really matter, the important point is that the thing will present itself differently in different situations. In other words, we are always already interpreting the world in accordance with rules to which we are committed.

upon it, sounds remarkably like the view of human subjectivity that Heidegger rejected. Rather his view insists that man is a being, "thrown" into the world, who is always already acting and coping in a pre-reflexive manner with things that have already acquired meaning. This is not at all dissimilar to the view propagated in this chapter, i.e. man as habitual rule-follower before he is a calculator[114]. It is also suggestive that these views were the direct result of his considerations regarding the influence of temporality on human life. In similar manner the conclusions reached in this manner were largely the result of considering the fact that human life occurs in time, as this implies computational costs. For these reasons I think that there are strong parallels between Simon's critique of the "rational choice" view of human agency and Heidegger's critique of overly rationalistic conceptions of human subjectivity.

10. Conclusion

The first two points made in this thesis that are crucial to the argument can both be seen as the inversion of a traditional view. In chapter 1 it was argued that ethical hedonism is logically prior to psychological hedonism. In chapter 2 it was argued that this enables a lot of the traditional objections concerning hedonism to be overcome. This chapter attempted to argue for a similar inversion. It argued, against the predominant view of the relation between rational choice and bounded rationality, that man is a rule-follower before he is a calculator.

This was done by firstly considering the conditions under which rule-following can become optimal, i.e. ignorance and information costs. It was then argued that the fact that human beings can incur computational costs implies that, all else being equal, rule-following is superior to calculation since it saves on computational costs. Several hypothetical examples concerning chess and other games were then used to argue that rule-following cannot be seen as an instance of a prior calculation. Rather calculation is, for a being with constrained cognitive capacities, something that can only happen once a context has been specified by the prior adoption, or inheritance, of rules.

---

[114] Calculation was portrayed as a secondary activity in way similar to Heidegger's portrayal of "theory" (calculation *par excellence*) is a derived activity *arising* from praxis ("rule-following"). See

This conclusion becomes even more pressing when one considers the fact that the mind has certain innate decision-rules. This means that the mind has been partly "programmed" by evolution, a notoriously "bounded" designer. This implies that the information-processing rules of the mind, while effective, will tend to be the type of thing that would make any self-respecting mathematician blush.

The above arguments lead to seeing rational choice theory as procedurally inaccurate, but as a useful tool of simplification. It was argued that this view presents an interesting angle from which to view Hobbesian contractarianism. It was also argued that there are striking parallels between the rational choice / bounded rationality dispute and Heidegger's critique of traditional philosophical conceptions of subjectivity.

What does this imply with reference to the main topic of this thesis? The first part of this thesis argued for cognitive, egoist, ethical hedonism. The question that this chapter started on is simply the question "How should an egoist hedonist act?". This is an instance of the more general question "What actions will maximise utility?".

The first main conclusion of this analysis has been reached. Normatively, this is that a utility-maximiser with the normal human capabilities should be a rule-following creature who occasionally calculates. Descriptively, it was shown that, as far as human beings go, we do not really have any other option available to us.

There are two main ways in which such a conclusion, if correct, will help to advance the argument in this thesis. This thesis argued for egoism based on the ontological nature of value. It argued that both ethical and psychological hedonism followed as the direct consequence of this argument. As such it has to, at some point, try and explain all human actions as expressions of egoism.

The view that all human action is egoistically motivated in some sense is not new in philosophy. The task of explaining particular actions as instances of egoism is a

---

Gelven (1970: 198-201) for a characterisation of how Heidegger views the relation between theory and practice.

daunting one, and has been undertaken with varying levels of success. There are great *prima facie* difficulties with making the acceptance of such a wildly counter-intuitive thesis a rational matter. Although egoists have stuck to their guns with vigour, it is probably fair to say that the balance of the evidence has weighed against them.

Hobbesian contractarianism represented a real break-through in making the argument for egoism less outlandish. Hobbes' argument that rational egoists would work together to establish society gave the idea of egoism more rational credibility than it previously had. Yet the evidence was by no means definitive. It is probably fair to say that only those who value the parsimoniousness of an explanation very greatly, or the very cynical, would be persuaded by Hobbes.

A similar breakthrough in rendering the argument for egoism less irrational came in this century with the study of iterated prisoner's dilemmas. This showed that cooperation can evolve without the need for a central authority (Hobbes' "sovereign"). This is a topic that will be discussed in chapter 6, where a host of other explanatory devices will also be employed in order to make the idea of egoism less strange.

I would contend that, if the argument in this chapter is correct, it helps to make the case for egoism similarly less implausible. As such it strengthens the case for egoism, though obviously not to the same degree as the two examples mentioned above. It helps the case for egoism in the following manner.

Traditionally authors have tried to show that, by starting with self-interested calculation, one can end up with a world in which the "moral phenomena" are accounted for. In this way they have tried to show that the following of ethical guidelines does not, in general, constitute a violation of egoism. Rather it is an instance of it, and can even serve as a normative base for it. This is the way the problem was conceived by Hobbes and others and, while they have had their supporters, this approach has not been generally accepted.

If the argument in this chapter is correct, then the above way of conceiving of the problem is incorrect. The problem is not to relate *calculating* egoism to altruistic

rules, but one of relating *rule-following* egoism to altruistic rules. In other words, one does not have to show that someone who is, in principle, a calculating egoist would be, in practice, a rule-following altruist. Rather one must show that a rule-following egoist would be, in practice, a rule-following altruist[115]. Or, at least, will exhibit a certain proportion of the actions we normally refer to as "altruistic".

Surely the task of showing that altruistic rules are an instance of egoist rules is somewhat easier than showing that altruistic rules are an instance of egoist calculations. Here there is one problem to be overcome, namely egoism to altruism, rather than two, namely egoism to altruism and calculations to rules.

This by no means clinches the case for egoism. It still needs to be shown that our intuitive expectation as to what egoist action looks like is wildly mistaken or untrustworthy. The argument for such a contention will be made in chapter 6. There, to be honest, everything but the kitchen sink will be thrown at altruism in order to turn it into egoism. But one point should already be clear. This is that, if the conclusions of this chapter are accepted, the idea of relating all human conduct to egoism has become slightly less strange, i.e. slightly more rational to accept.

There is another important implication of this chapter that needs to be mentioned. This one pertains to the version of hedonist egoism that is being defended here. It was stated at the start of this chapter that egoist hedonism isn't nearly as bad as it sounds, nor as anathema to traditional theories of ethics as is commonly supposed. Rather it is the case that the nature of the correct applications of these doctrines are counter-intuitive. The way human decision-making was portrayed in this chapter must already give the reader some indication of why this is the case.

I would venture that a large part of what is odious in our uncritical conception of hedonist, egoist action stems from the following view. An egoist hedonist is someone who considers the different possibilities, calculates the one that will give greatest pleasure to herself, and acts upon it. This is an extreme "rational actor" view of

[115] Strictly speaking, one must show that a *rule-following* egoist will be a *rule-obeying* altruist. Since a rule that is followed is also obeyed, showing that a rule-following egoist would be rule-following altruist is a way of *also* achieving this primary objective.

human agency, one that this chapter argued to be false. The form of hedonist egoism that is defended in this thesis is a much weaker one. It states that people are rule-following beings, i.e. attempt to achieve a bewildering array of objectives other than egoist hedonist ones. In fact, a human being is never a hedonist egoist in the strong sense portrayed above. A human being can only calculate which option will produce the most personal pleasure within a very restrictive context. This context will be the result of rules that determine which options are pre-selected and which rules of evidence are to be heeded. In other words, egoist pleasure can only function as a choice-criterion within a wider context of rules (or "norms", or "values"), which are being followed.

Is there still any sense in which people can be said to be egoist hedonists? I would contend that, despite the fact that egoist hedonist concerns only rarely influence our choices explicitly, and then only partially, that there is. Any given utility-function that is best implemented by satisficing will, over time, be the deciding factor in determining the nature of the satisficing criteria. In a similar manner the degree to which a *learned* rule implements hedonist, egoist concerns will always either reduce or increase the chance of it being retained or rejected. Any other principle can, if the argument in chapters 1 and 2 are correct, only function as a criterion for the rejection or retention of a rule contingently. In this way the logic of the evolution of learned rules will tend to favour those rules which best implement egoist, hedonist concerns. In this extremely weak sense pleasure is still "behind it all", and in this sense people should be, and are, egoist hedonists.

The next chapter will examine one specific "rule" that we are capable of following. This is the rule that tells us to adopt "truth" as a norm of inquiry. As such it will serve as a demonstration of how an ethical principle can be derived from the imperative to optimise utility, here conceived egoistically and hedonistically.

Chapter 5: Cognitive Ethics – The Value of Truth

1. Introduction

In chapter 1 and 2 an argument was presented for the contention that ethical imperatives can be true or false in the same sense as "factual statements" are taken to be true or false. In the previous chapter it was argued that, under certain conditions, these imperatives, expressed as rules, maximise utility. This chapter will attempt to combine these two insights and inquire as to the content of one of these rules. In doing so it is a first attempt to determine the *type* of ethical imperatives that can be derived from the meta-ethical considerations defended at the start of this thesis. It will also serve as a demonstration of the methodology that is appropriate for establishing that an ethical rule is true or false.

In the previous chapter man was painted as a rule-follower that occasionally, and in a specific context, calculates. Given that man has certain criteria by which to judge beliefs, certain rules that assess the evidence, and a limited set of possible beliefs, man will judge these beliefs against the criteria, i.e. calculate. This is, however, a far cry from calculation "all the way down". Again, as in all matters, the criteria, evidential rules and decision as to which options to include will be a result of rules. It is only within the context afforded by an adherence to these rules that calculation is possible.

This chapter will attempt to determine the criteria by which beliefs should be judged. It will ask what rule a rational actor with the normal amount of computational capacity and informational constraints will use to judge beliefs, if indeed it should use a rule at all. It will be argued that a rational actor should indeed use a rule, and that this rule will instruct him to use "truth" as criterion for belief-formation. This rule is simple enough to recommend it to a satisficer with the normal human capabilities. In other words I wish to defend the claim that "truth" should be a norm of belief-formation. If the argument presented in chapters 1 and 2 is correct, then this claim is true or false according to the same logic that "snow is white" is true or false.

The chapter will begin by defining the idea of truth. It will then argue that this should be the ultimate criterion of belief-formation. The chapter will end by taking advantage of the perspective developed within to look at some traditional ethical doctrines. These share the peculiar quality that they rest on the normative assumption that truth is valuable. Yet the authors treat it as obvious enough not to try and justify this by argument. It seems possible that they do not fully realise they have made a value-judgement, since the works under discussion all dismiss "value-judgements" from the realm of rational discussion.

2. Definitions

2.1 "Truth"

The exact definition of "truth" has been a matter of frequent dispute in philosophy. Fortunately such disputes can be avoided in this chapter. "Truth" will be defined in terms of only one property. This property will surely be granted to belong to "true" statements, regardless of what one's full definition of truth is.

This property is based on the following: if one's expectations with regards to the *empirical consequence* of believing *a specific statement* are upset, this means that the statements giving rise to these expectations were false. If one has expectations with regards to the empirical consequence of true statements, these expectations cannot be upset in the same radical manner. Simply put, true statements do not lead to surprises in the same way that false statements do.

This forms the base for the characterisation of "truth" that will be used in this chapter. Most of our actions are based on a mixture of true and false beliefs. If our expectations are upset, or the consequences of a specific action is surprising, then this is normally ascribed to false beliefs that the actor had. We will not call any statement to which we can ascribe such a "surprising consequence" a true belief. Hence "truth" can, for the purposes of this chapter, be defined in the following way: a "true"

statement is one that will not lead to surprises, in the sense of upset expectations[116]. The assertion that true statements have this property would seem to be consistent with just about any conceivable definition of truth.

As a brief example to illustrate the above, consider the following: A walks to the bank at 16h00 to deposit some money. A has the false belief that the bank is still open at 16h00. A's expectations will be upset when she is faced with a closed door at the bank. This surprising consequence can be ascribed to the false belief that the bank is open at 16h00. This would not have been the case if A had acted on the true belief that the bank is still open at 15h00. The consequences of acting on such a true belief would have been unsurprising, i.e. A would have passed through the door of the bank and been served.

2.2 "Holding a belief"

In this chapter I wish to argue that people should hold beliefs that are true. It might be objected that to "hold a belief" and to "hold a belief to be true" is the exact same thing. This would reduce the argument in this chapter to a confused tautology.

For the above reason a behaviourist definition of "holding a belief" will be used in this chapter. On this definition, "to hold a belief", is to act in a certain way when the belief becomes relevant to action. In this way believing that one should stop at red lights can be defined as a disposition to stop at red lights.

The range of beliefs being dealt with in this chapter will also be restricted in order to simplify the argument. Only beliefs that are relevant to action will be discussed. Hence any belief, of which I can believe the negation without it changing my actions, is excluded from the discussion.

Note that the question asked in the chapter concerns the situation *where it is a given* that someone will have beliefs about something. The only remaining issue is with

[116] Whether there is a matter of fact as to which belief a "surprise" can be specifically ascribed to, or whether a degree of convention enters the picture is not relevant. What matters is that we will not call the statement we "blame" for the surprise "true".

regards to the standard used to select these beliefs. The question whether someone should have beliefs *at all*, and about what, will not be directly discussed. Hence tricky issues regarding situations where uncertainty is part of the "thrill" or mystery of the situation will be avoided.

3. The Value of Truth – The True and the Good

The main question being asked in this chapter is "What type of beliefs should an Egoist hold?". The deceptively simple answer must be that an Egoist should hold beliefs that will lead to the highest value of experience over time. Hence it must be asked "What type of belief will lead to the highest value of experience over time?". It is with reference to this question that the value of "true belief" will be assessed. "Beliefs that lead to the highest value of experience over time" will be referred to as "good beliefs", and it will be asked whether "true beliefs" are necessarily "good beliefs".

Are "true beliefs" always "good beliefs"? The answer to this must be an unambiguous "no". Many situations can be constructed where someone would have been much better off if his beliefs had been false. Imagine someone going to the cinema as the result of a false belief that a certain film is showing. Now imagine that the person discovers his mistake, turns to leave and is surprised to find an old friend standing behind him. This chance meeting leads to them renewing their friendship, and proves very beneficial to both parties. Here there is a clear case of an action based on false belief that is most definitely advantageous to the person. It is quite possible that renewing the friendship might not have occurred in a possible world where the person had not acted on this false belief.

Is the opposite then true, i.e. are "true beliefs" always "bad beliefs"? Again the answer must be an unambiguous "no". A true belief that the brakes on one's car need to be replaced before it will be safe to resume driving can leave you considerably better off than if you had incorrectly believed the opposite.

"True beliefs" can be "good" sometimes, "bad" at other times. In the vast majority of cases, however, "true beliefs" will also be "good beliefs". To justify this statement it is

sufficient to consider the vast number of beliefs that are relevant to most human conduct. Consider the number of beliefs that are necessary if someone is to, for example, build a house. Any mistake regarding the needed materials, method of construction, relevant engineering principles, etc. could result in the project being unsuccessful. The idea of successfully building a house if most of our beliefs about the relevant actions are false is almost nonsensical. A well-built house is not something that can be delivered by chance alone. Hence, while it is sometimes true that a "false belief" can be a good belief, "false belief" is always unlikely to be "good belief".

The argument thus far can be summarised as follows: "The True" will tend to coincide with "the Good", but will not always coincide thus. This leads to viewing "truth" as the fundamental norm of inquiry, as will be discussed below.

4. The Value of Truth - "Truth" as a Norm of Inquiry

4.1 An Argument for Truth as Norm of Inquiry

It often seems "obvious" or, indeed, tautological to state that "truth" should be the norm of inquiry. It is not, however, as obvious as it is often taken to be. Beliefs can have many different qualities apart from their truth-values. We can distinguish between beliefs based on whether they are true or false, useful or useless, whether they are held by another or not, whether they can be expressed in ten words or less in a given language or not, etc. It is by no means obvious that only true beliefs should be held, for it is logically possible to adjudicate between beliefs based on an infinite number of other criteria.

I now wish to argue that we should, once we have judged it necessary to have beliefs concerning a given subject matter, use "truth" as our explicit guide in choosing which beliefs to hold. I will proceed by stating a fairly straightforward, and, indeed, popular argument for the instrumental value of holding true beliefs[117]. This argument, however, stands in need of some qualification because of the human capacity for self-

[117] For example, Joyce (2001: 178-179) gives an argument similar to the one given below, and interprets James and Peirce as having given versions of the same argument.

deception. It will then be argued that the capacity for self-deception has some limitations that allows a slightly weaker version of the argument for viewing "truth" as a norm of inquiry to succeed.

The simple argument, in terms of the method of analysis outlined in chapter 4, runs as follows: An Egoist should aim at holding good beliefs. To do so she needs a criterion to distinguish "good beliefs" from "bad beliefs". How can it be judged whether a belief is "good"? This is a matter of determining the consequences of holding a belief. An Ideal Egoist would investigate the consequences of holding a belief before deciding whether a belief should be held. It is here where we run up against the limits of our knowledge and computational abilities. None of us have the ability to predict the future with any certainty.

The simple argument that I wish to present is designed to reach the conclusion that people should habitually use "truth" as a criterion by which to decide whether a belief should be held or not. In terms of the method discussed in chapter 4 it needs to be shown that this can be formulated as a rule that meets the three criteria for rule-following outlined in the previous chapter. Such an argument would normally start by showing that it is almost impossible to know when it is beneficial to hold a false belief. In the case of "truth", however, it is not necessary to make this argument in order to show the necessity of rule following. In the case of truth it is not *almost* impossible to know when a false belief is better than a true belief, but *always* impossible.

When the case of belief is considered an extra factor enters the equation. This is the fact that a positive outcome based on a false belief is always an unforeseen outcome. For if such an outcome was foreseen it means I had a "true belief" about it, and then I could have chosen this outcome based on this "true belief". As such the false belief would no longer be causally necessary to bring about the positive outcome. It can be brought about by choice based on true belief.

Hence, when it comes to the matter of "true belief", an Egoist should follow a rule, rather than judge cases individually. This can be made fully evident by using the criteria for rule following stipulated in the previous chapter:

1. A situation recurs that is alike in a certain respect: An actor has to choose whether to hold true or false beliefs.
2. Consistently performing one action is clearly preferable to consistently performing its opposite: In the vast majority of cases true belief is also good belief.
3. The actor has no way of knowing when it is better to perform the opposite of the action, and/or the cost of obtaining this information is prohibitive: The case where a false belief is preferable cannot be foreseen, since foreseeability presupposes a true belief about the situation which could be used to bring about the advantageous state independently of the false belief.

From the above it should be clear that, since an Egoist should aim at good belief, this implies that she should aim at true belief. If inquiry is defined as an attempt to determine good belief, then it follows that "truth" should be a norm guiding this process. Or, to use the metaphor explained in the previous chapter, "good belief" is the gold coin to be found at the bottom of most "true beliefs".

4.2 Objection Based on "Self-deception"

The above argument stands in need of some qualification. Above it was stated that any positive consequence based on a false belief could also have been achieved by some other true belief. And, if I know that the false belief has a certain positive outcome, then I also know the true belief that will guarantee this outcome. Hence I no longer need the false belief, and can disregard it. Mostly this is the case; if I know that the false belief that a film is showing at eight will, by blind luck, cause me to meet a long-lost friend, then I also have the true belief regarding where I can meet the long lost friend. This implies that the positive outcome need no longer depend on the false belief.

There is, however, a class of actions where there is no true belief that can bring about the same beneficial outcome. Imagine I am nervous about presenting a paper, but the false belief that the audience likes me gives me the confidence to do a good job. Here the false belief is causally necessary in order to bring about a positive outcome. This positive outcome cannot necessarily be brought about by a true belief. The fluency of

one's speech, the confidence one projects, etc. are not, sadly, under one's full conscious control. I might well try and guess what I would do if I had the false belief that the audience likes me, and then try and imitate this hypothetical me. But it will surely be uncontroversial to assert that, when it comes to confidence, there is often no substitute for the real thing, no matter how misguided.

In the above example it would be in the person's interests to deceive himself about the nature of the audience's attitude towards him. And, to the degree that this is psychologically possible, it would seem an inescapable conclusion in terms of the main argument of this thesis that such a course of action is to be recommended.

4.3 Relevance of Self-deception

Self-deception is a phenomenon that has been of interest to philosophers, psychologists and others for quite some time. The phenomenon is well documented[118], and, in terms of the analysis made in chapter 4, its existence need not be particularly surprising. If a great number of the information-processing mechanisms in the modular mind are the result of evolution, then there is no reason to suspect that these mechanisms were selected solely with reference to their ability to discover truth. Rather they were selected based on their contribution to the reproductive fitness of the individual. And if, in cases like assessing the confidence one should have in oneself, it would be advantageous for these rules to be set so as to be slightly biased towards an overly positive evaluation of oneself, then such rules would be selected.

The above type of situation, namely one where I can sometimes not bring about a positive result based on a false belief in any other way, has frequently been cited as providing an evolutionary rationale for the existence of self-deception. For instance, Trivers (1981)[119] pointed out that being able to deceive others can be highly adaptive. But people are not perfect liars, for whatever reason people do tend to unintentionally signal the fact that they are lying. This would give rise to an evolutionary pressure in

[118] Some interesting case studies can be found in a collection, entitled *The Multiple Self*, edited by Elster (1985).
[119] For a discussion, see Badcock (2000: 132-134).

favour of people who are very good at detecting lying, and thereby negate the advantage offered by deception. However, if one does not know that one is lying, i.e. if self-deception is possible, then the usual methods of lie-detection would fail. In other words, "the organism is selected to become unconscious of some of its deception" (1981: 35)[120].

What does this imply for the argument made above to the effect that a rational egoist should be a truth-seeker? Firstly, it implies that any given person might hold a lot of beliefs that aren't true, but are useful at maximising reproductive fitness. And, since reproductive fitness and utility maximisation are often overlapping goals it might not necessarily be a good idea to try and change them.

For the above reason I will restrict my claims for the virtue of truth to beliefs that we do not yet hold. In other words I wish to restrict the above "simple argument" only to cases where I have ascertained that I need beliefs concerning a topic on which I do not yet have any detailed beliefs. This is not to argue that I should not, perhaps, try and correct beliefs that I already hold. Rather I will exclude such cases from this discussion.

This analysis will also be restricted to cases where I have to consciously decide whether a given belief should be held. It will surely be uncontroversial to assert that I am somewhat constrained when it comes to self-deception. I am pretty sure that, no matter what I do, I cannot now *consciously* decide to believe that, for example, I owned a Porsche when I was thirteen years old. In other words I am claiming that it is *psychologically impossible or prohibitively difficult* to believe that I did own a Porsche, realise that it would make me happier if I did believe this, and therefore to decide to believe this.

If this is the case, then a qualified version of the "simple argument" given above can be defended. Consider a situation where I have already decided to acquire beliefs about a topic, regarding which I now have none. There will be some false beliefs that

[120] Such a logic seems well-suited as a partial explanation for "hindsight-bias", i.e. the documented fact that people tend to overestimate the accuracy of their past predictions. For another justification based

would have outcomes more positive than the true beliefs. These outcomes can now be distinguished into two groups based on whether these outcomes could also be brought about by true belief, or not. If the positive outcome in question could also be brought about by true belief, then this presents no obstacle to arguing that one should always aim at true belief.

If the outcome could not have been brought about by true belief, then this still does not amount to an argument against the notion that, when consciously weighing belief, one should aim for true belief. This is because of the claim that our capacity for self-deception is constrained in such a way as to make a *conscious* choice to believe something you know to be false impossible.

The claim that truth should be a norm of inquiry can now be defended in the following, weaker form: When attempting to *consciously* acquire beliefs concerning a given topic, try to achieve only true beliefs[121]. For the rest of this thesis I will refer to the claim made here as the idea that truth should be a norm of inquiry.

This renders the main objective of this chapter achieved. This objective was to illustrate the methodology of a cognitive Egoist ethics with reference to a concrete example. It has been argued that the rule "If you have to consciously decide which beliefs to hold, then you *should* attempt to hold true beliefs" is true. The fundamental point of the argument was the part that was always implicit. This is that, due to the arguments made in chapter 1 and chapter 2, the above rule can be judged as true (or false) in the same sense that "Snow is white" can be judged to be true or false.

The next chapter will concern the form of interaction that is egoistically rational. First I wish to take advantage of the view afforded by the preceding discussion to look at certain ethical doctrines. These seem to depend on the value of truth and yet do not explicitly admit this.

---

on computational savings, one that need not contradict the view proposed here but can complement it, see Hoffrage and Hertwig (1999: 191-208).

[121] Note that I am not claiming that the opposite is the case concerning already existing beliefs, beliefs irrelevant to action, etc. Rather I am simply excluding them from this discussion.

5. The Moral Condemnation of the Denial of Truth

Certain ethical doctrines, which will be looked at below, can be cast in the following mold:

1. A certain fact is asserted to be "manifest truth".
2. A certain action is then asserted as being an implicit denial of this "manifest truth".
3. This denial of a manifest truth is used as a basis for moral condemnation.

What is missing from the above reasoning? It needs to explain what is so bad about denying, or hiding from, a manifest truth. For, even if someone concedes that a certain action is tantamount to the implicit denial of a manifest fact, the person can still maintain that there is no reason to suppose the denial of such a fact to be morally wrong. In short, what is missing is an explanation of why truth is valuable, and why manifest truth should be admitted. This explanation is what this chapter has attempted to provide. Arguments of the above form are incomplete, since they fail to state why actions should be guided by true belief.

For a famous version of this argument, consider Sartrian "bad faith", as stated according to the form outlined above (Sartre, 1981: 626):

1. Moral principles are not independent of subjectivity.
2. To act in the "spirit of seriousness" is an attempt to hide this fact from oneself; it is to be in bad faith[122].
3. "The principal result of existential psychoanalysis must be to make us repudiate the *spirit of seriousness*" (1981: 626).

The argument is as follows: To act in a spirit of seriousness is to act as if there are values independently of subjectivity. Sartre denies that there are any such values. From this he draws the conclusion that one should not ("must...repudiate") act in such a manner.

---

[122] "Man pursues Being blindly by hiding from himself the free project which is this pursuit" (1981: 626).

Such an argument is incomplete unless it has been explained why one *should* not attempt to deny such facts to oneself, and unless this argument contains a self-justifying element[123]. What is lacking is an explanation of the value of truth, i.e. an explanation of why one *must* reject what isn't true. Without such an explanation it is unclear how Sartre can generate the ethical "must" in the above argument.

Wittgenstein's most famous pronouncement operates according to the same logic:

1. There are no ethical facts[124].
2. To speak about ethics can only be possible in denial of this fact.
3. "Whereof one cannot speak, thereof one must be silent" (1983: 189)[125].

Above the "fact" that there are no ethical facts is used to generate the imperative that one "must be silent". Presumably this self-undermining pronouncement was Wittgenstein's conscious demonstration of the hopelessness of ethics if we accept his premises. It does, however show how natural it is to assume that truth is valuable. For, even after "proving" – conclusively to his mind – that there can be no ethical pronouncements, he cannot quite refrain from justifying his silence on matters of ethics by including one self-undermining imperative.

Further examples could be enumerated without really gaining any clarity with reference to the main object of this thesis, i.e. meta-ethics[126]. One extract from Bentham needs to be included, however, to show how far the above type of reasoning can be stretched.

> We have one philosopher (Woolaston), who says, there is no harm in any thing in the world but in telling a lie: and that if, for example, you were to murder your own father, this would only be a particular way of saying, he

[123] Sartre would deny the possibility of such a self-justifying element: "Ontology itself cannot formulate ethical precepts" (Sartre, 1981: xiii).
[124] "The good is outside the space of facts" (Wittgenstein, 1984: 3).
[125] Note that "cannot" here means "cannot sensibly", not "cannot" in the absolute sense.
[126] Cilliers (1998: 139) makes an argument of the same type: "To fall back on universal principles is to deny the complexity of the social system we live in, and can therefore never be just." Here the implicit

> was not your father. Of course, when this philosopher sees any thing that he does not like, he says, it is a particular way of telling a lie (Bentham, 1982: 27).

6. Conclusion

This chapter tried to determine a specific instance of rule-based behaviour that can be recommended to an Egoist. It was found that "truth" should be employed as a norm of inquiry. In other words, the rule that, when consciously trying to acquire beliefs, one should aim at truth, was argued to be a valid ethical rule. This is such a basic intuition for many philosophers that it is sometimes used to generate ethical conclusions by authors who claim to have given up on the idea of ethical conclusions.

The next chapter will concern the objection most easily made against egoism. This is the objection that the existence and desirability of altruism amounts to a refutation of egoism, both as a descriptive and normative doctrine. The next chapter will try and argue that our intuitions as to what egoist action should look like are radically mistaken.

---

denial of the complexity of the social system is a "manifest fact" that is denied if one falls back on illusory principles. This denial is false, and hence concluded to be unethical.

Chapter 6: The Possibility of "Altruism"

1. Introduction

The first chapter of this thesis presented an argument for cognitive, ethical, egoist hedonism. It was claimed that there is an element of our experience, often referred to as pleasure, that can count as the ultimate referent of axiological claims. It was also claimed that this ethical hedonism implies a weak form of psychological hedonism.

This argument, by itself, is not yet sufficient to generate any specific claims as to how an ethical hedonist should act. Such claims are only possible once an account of the decision-making mechanisms used to generate action has been given. This account was given in chapter 3. There an argument was presented for the contention that, given the cognitive constraints and design history of the human mind, human beings can only ever be boundedly rational. In other words, human beings can only ever be rule-followers, not calculators.

This then raises the question as to which rules an ethical hedonist should follow. The previous chapter gave a defence for a qualified version of the rule that truth should be a norm of inquiry.

This argument marked the end of the constructive part of this thesis. This thesis is an attempt to outline a coherent meta-ethical position that allows for a cognitive ethics. It was explained why people should follow moral rules, why these rules can be viewed as objectively true or false, and an example was given of such a rule. This chapter will attempt to defend this general outline of meta-ethics against the most counter-intuitive implication that it has. Or, at least, against the most counter-intuitive implication that it would *seem* to have.

This implication regards that which most people would consider to be the very epitome of ethical action, namely altruistic action. It would seem that there is nothing in this thesis that provides for a commendation of altruism, or even an explanation of its existence.

Can anything in this thesis be used so as to justify actions that we would normally regard as altruistic? It should firstly be noted that a defence of altruism need not, as a matter of logical necessity, be a part of any treatment of ethics. It is not *logically impossible* to suppose that there exists a valid meta-ethical system that condemns altruism, or, for that matter, violates our main ethical intuitions in any manner conceivable. Such a violation would merely be a counter-intuitive implication on a par with the counter-intuitive implications to be found in just about any other academic discipline. If it can be shown that ethical rules have truth-conditions in the same sense that assertions in physics have truth-conditions, then implications with regards to ethics that condemn altruism are logically no more strange than implications in physics that profoundly violate our intuitions with regards to space and time.

The above reasoning does not, however, justify disregarding the topic of altruism as merely a counter-intuitive implication of the meta-ethical theory. For, although the existence of altruism cannot logically serve as a knockdown argument against such a theory, it does raise several questions. If indeed, a meta-ethical theory condemns altruism then it needs to explain why our moral intuitions are so flagrantly false and why these falsehoods are so widespread[127]. This responsibility falls to any theory that upsets long-held beliefs.

In physics an explanation of why our intuitions are false is relatively easy. Our view of space and time is quite adequate to deal with the everyday world. The phenomena that call this view into question only arise when distances and magnitudes are encountered that are of no relevance to what we normally describe as everyday life.

---

[127] Certain basic ethical rules that justify altruistic action seem to almost be cultural universals. Consider the rule "Do unto others as you want done unto yourself". The following religious documents all include a version of this rule as a basic ethical teaching (taken from Harris et al (1995: 167)):

Christian version: "Treat others as you would like them to treat you". (Luke 6: 31, New English Bible.)
Hindu version: "Let not any man do unto another any act that he wisheth not done to himself by others, knowing it to be painful to himself" (*Mahabharata*, Shanti Parva, cclx.21).
Confucian version: "Do not do to others what you would not want them to do to you" (*Analects*, book xii, #2).
Buddhist version: Hurt not others with that which pains yourself" (*Udanavarga*, v. 18).
Jewish version: What is hateful to yourself do not do to your fellow man. That is the whole of the torah" (*Babylonian Talmud*, Shabbath 31a) (continued next page).

No such answer, however, suggest itself with regards to the position of altruism in ethics.

This chapter will present the evidence for the contention that our view regarding altruistic action is *not* radically false. In other words, it will present the case for supposing that an egoist ethics could still recommend a lot of actions that we would normally consider to be "altruistic".

At this point it is necessary to clarify exactly what will be argued in this chapter. It will be claimed that our intuitions regarding the actions that an egoist hedonist should commit are untrustworthy. It will not, however, be claimed that they are definitely false. In order to claim that a large number of actions that an egoist hedonist should commit are of the type that we would normally call altruistic, a very complex argument needs to be made. One would have to present an argument regarding the preferences of humanity, the decision mechanisms that implement these preferences and the game theoretical structure of human interaction. Such an argument would presuppose great expertise in evolutionary biology, very specific knowledge regarding human cognitive mechanisms, and an expert knowledge of psychology, sociology and game theory. It is certain that there is no one person who has all the knowledge that is necessary for such an argument. In fact, I think it is safe to say that even the sub-parts of such an argument, i.e. the specific parts relating to one of the specific disciplines mentioned above, are beyond the knowledge that humanity currently possesses.

For the above reason the best that can currently be done is to give an account of some of the different ways in which an egoist ethics could possibly be made consistent with the existence of altruism. This will not amount to a conclusive argument for altruistic action, but will be an attempt to demonstrate that it is far from certain that most of the actions that we normally call altruistic constitute an exception to egoism, rather than a manifestation thereof. In other words, this argument will try to undermine the justifiability of "altruism" *qua* objection to the idea of an egoist ethics.

Muslim version: "No man is a true believer unless he desires for his brother that which he desires for himself" (*Hadith*, *Muslim*, *imam*, 71-2).

2. Outline of this Chapter

There are four main ways of trying to show that an apparently altruistic act does not constitute an objection to the idea of an egoist ethics. Firstly, it can be shown that the altruistic act itself is caused by the altruistic *preferences* of the actor. Or, in terms of the view argued for in this thesis, it can be shown that the actor derives pleasure from the apparently altruistic act. Secondly, it can be shown to be the result of the *game theoretical structure* of human interaction. In other words, it can be claimed that the situation in which the actor finds herself is misleading to our intuitions in such a way that an apparently altruistic act can actually be shown to be self-interested. Thirdly, it can be argued that the apparently altruistic act, while not directly caused by the preferences of the actor, is the direct result of the bounded rationality of the actor. In other words, it can be argued that the act is the result of a given set of preferences, in conjunction with our *cognitive constraints*. The fourth way of turning apparent altruism into egoism is by arguing that the action rests on some *factual mistake*. In other words it can be argued that an apparently altruistic act rests on some error made by the actor, and that this error is a causally necessary condition for the occurrence of the act.

This chapter will proceed by explaining each of the above four methods in turn, and present evidence for supposing them to be relevant to human action.

3. The "Altruistic Preferences" Defence

3.1 Introduction

The easiest way of making apparently altruistic action consistent with egoist hedonism is to claim that the actor derived pleasure from performing the actions. This type of defence of maximising behaviour in the face of apparent altruism is most familiar to economists. Any action that does not serve the obvious self-interest of the actor, and is not the result of any factual error is normally taken to reveal something about the utility function of the actor. The uses and pitfalls of this approach have been discussed in chapter 3. Since hedonist egoism is a form of maximising behaviour, and

specific preferences can be reinterpreted as causes of pleasure (as explained in chapter 1), the following discussion also directly concerns economics.

Preferences affect behaviour, and behaviour affects the reproductive fitness of the individual. This means that preferences can be viewed as a proximate mechanism whereby evolution influences action. This presents a standard for judging the likelihood that a certain preference commonly exists. If it is claimed that a certain preference is commonly found among human beings, then an account of how such a preference provides an evolutionary advantage is *prima facie* evidence that the preference does exist. This type of reasoning is not uncommon; it is precisely why we commonly assume that most people care greatly about self-preservation and sex. The opposite is the case with preferences that detract from evolutionary fitness. If someone claims that most people throughout history had a taste for fatally poisonous food we would not seriously entertain such a claim. Any such behavioural trait would be too strongly selected against for such a claim to be at all worth considering.

3.2 Evolutionary Altruism and Psychological Altruism

The discussion to follow depends critically on what is meant by "altruism". Since preferences will be portrayed as having been shaped by evolution, it is necessary to draw a distinction between the evolutionary concept of altruism, and the everyday or psychological concept of altruism.

Evolutionary altruism is commonly defined as action that would decrease the fitness of an individual relative to the group that he is part of, while increasing the fitness of another individual in the group[128]. The evolutionary concept of altruism radically differs from the psychological notion of altruism. It only relates to the consequences of behaviour, and says nothing of the proximate causes of behaviour. Hence an "actor" can be an altruist in the evolutionary sense without having a mind, without having any intentions, and even despite having selfish intentions (Sober, 1998: 460-462).

---

[128] I am here following the treatment of "evolutionary altruism" and "vernacular altruism" ("psychological altruism") to be found in Sober (1998: 459-487) and used by Sober and Wilson (1998) in their definitive treatment of the modern theory of group selection.

In contrast to the evolutionary notion of altruism, the psychological notion only applies to creatures with minds, and therefore intentions. To judge an act as psychologically altruistic is to say that it was based on an ultimate desire that had irreducibly other-directed propositional content, and aimed at the welfare of another (Sober and Wilson, 1998: 199-231). An altruistic desire needs to be *ultimate* in the sense that it cannot be derived from a more fundamental egoist desire. It must have *propositional content that is irreducibly other-directed* in the sense that the proposition expressing the desire must include reference to someone else. In other words, it must be something like "I want Jones to have the apple". It must also concern the *welfare* of another so as to exclude cases of spite or malicious intent. For example, the desire "I want Jones to suffer social embarrassment", while other-directed, would not count as altruistic.

One matter needs to be clarified before proceeding. In chapter 1 it was argued that "preferences" or "desires" cannot be the ultimate causes of action. It was claimed that they fail to be self-justifying in the sense that a cause of action needs to be in order to be ultimate. Yet I will continue to speak of preferences, since it is in terms of preferences and desires that the academic debate relevant to the argument being made here has mostly been conducted. In accordance with chapter 1 all reference to preferences can be interpreted as talk of the causes of pleasure. In other words, to say that I have the altruistic desire that Jones should have the glass of water can be interpreted as saying that the possible world in which I know that Jones has the glass of water has greater value to me than the possible world in which he does not. Hence, when it is claimed below that a desire that contributes to evolutionary fitness is altruistic, this should be reinterpreted as meaning that an egoist hedonist derives pleasure from certain states of the world because they are to the benefit of others.

3.3 Kin-selection

In terms of the above definition, the task of this chapter can be reformulated as trying to explain why a hedonist egoist will have some ultimate desires that are irreducibly other-directed, and concern the welfare of the other that it is directed at. The first source of altruistic desires is an idea known as kin-selection.

In 1963 WD Hamilton published a paper entitled "The Evolution of Altruistic Behaviour" that shed great light on problems in evolutionary biology that had previously appeared unsolvable. Essentially it is an elegant way of explaining why altruistic actions towards related individuals constitutes an instance of survival of the fittest, and not a refutation thereof. Hamilton adopted the "gene's point of view"[129] and compared the fitness of two hypothetical genes. One of these cause altruistic behaviour towards genetically related individuals, while the other does not. If the disadvantage to the carrier of the altruistic gene is small enough not to outweigh the advantage to the helped individual (weighted by the degree to which they are related), then the altruistic gene will be more likely to leave copies of itself in the next generation.

Here the increase in the relative fitness of the gene will cause it to be selected, *despite the fact that it decreases the fitness of its owner*. In other words, any gene that causes helping behaviour to related individuals will have greater fitness, all else being equal, than genes that do not, as long as the cost of helping does not outweigh the benefit to the helped, as weighted by relatedness[130].

3.4 Kin-selection and Altruistic Preferences

Kin-selection is a powerful idea that gives an elegant explanation of why beings that care for their genetic "families" will tend to be favoured by evolutionary processes over beings who do not. As such it is a natural starting place for showing how altruistic preferences can evolve.

Note that the proximate mechanism whereby kin-selection is facilitated need not *necessarily* be psychologically altruistic, in the sense of being other-directed. A being that was born with the desire to optimise the inclusive fitness of his genes would display all the helping behaviour needed for kin-selection to operate, without any altruistic desires being required. No one, however, will seriously contend that a being with such an ultimate desire has ever existed. Rather the proximate mechanisms

[129] This idea of adopting the "gene's point of view" formed the basis of "selfish gene theory", as popularised in Richard Dawkins' *The Selfish Gene* (1976).

whereby evolution would implement such an agenda is likely to consist of one or more other-directed desires that do a passable job of carrying out a kin-selection agenda (Joyce, 2001: 136). This could be anything from a general desire for the welfare of a sibling to more specific desires like wanting siblings to be healthy, live a long time, have families, etc.

Also note that the actual desire is unlikely to make any reference to actual genetic relatedness. Rather the environmental cue that the desire is directed at is likely to be something that is merely correlated with genetic relatedness. For example, a general concern for people that one is brought up with could be used to implement a kin-selectionist agenda (Joyce, 2001: 136).

Kin selection presents a first method for arguing against the idea that egoists will be people who only commit the type of acts that are normally labeled as "selfish". Rather people can, when genetically related individuals are concerned, be expected to have a set of preferences that are altruistic. In chapter 1 it was argued that an altruistic desire must be reinterpreted as an egoist concern for pleasure deriving from the object of the altruistic desire being realised. The argument presented there, in conjunction with the idea of kin-selection, implies that a general concern for parents, siblings, cousins, etc., cannot be used to object to a characterisation of people as egoist.

3.5 Group Selection and the Problem of Evolutionary Altruism

Kin selection can help to explain exceptions to individual fitness maximisation because the individual is not the unit of selection. Rather the gene is the unit of selection, and this has implications for the behaviour of the individuals[131]. A similar

---

[130] A being that does this is said to maximise "inclusive fitness", where "inclusive" is understood in opposition to "fitness" as classically conceived.

[131] There is a large literature on what should count as the fundamental unit of selection in evolution. Some argue for the gene as fundamental unit of selection, some for the individual and some argue for multi-level selection. In this thesis I will mostly use the multi-level view argued for in Sober and Wilson (1998). This should not be taken as an important commitment of this thesis, since all the substantial claims made in this chapter can be explicated in terms of other views. For example, even the group selection argument can be rewritten in terms of selfish gene theory, should one wish to do so. Alternatively, kin selection could be rewritten to appear to be an instance of group selection (Sober and Wilson, 1998: 55-100). Hence the partial use of Sober and Wilson's conceptual frame is based on ease of use, rather than some substantial commitment. For a good introduction to the unit of selection

situation obtains with regards to "group selection", where the group, and not the individual, is the unit of selection.

Kin selection is one of the recognised triumphs of twentieth century evolutionary biology. This is not the case with group selection. A layman who is used to thinking of evolution in terms of "survival of the species" would be surprised to discover that "survival of the species" is an explanation that biologists view with great suspicion. But reports of the death of group selection are greatly exaggerated. Long viewed as a last resort for the softhearted (and possibly soft-headed), it has recently gained in credibility as a method of explaining evolutionary change.

The apparent impossibility of group selection is a direct result of the apparent impossibility of evolutionary altruism. Evolutionary altruism was defined earlier as the behaviour that lowers the fitness of an individual relative to that of the group. At first sight it might appear that this constitutes the very paradigm of behaviour that could not possibly evolve. One might try to defend group selection by supposing that certain behavioural traits, while decreasing the relative fitness of the individual within the group, could still increase the fitness of the group as such. This does not, however, seem to matter. For any behaviour that decreases the relative fitness of the individual within the group will be selected against, no matter how much it helps the group. All individuals in the group will be benefited by behaviour that helps the group, whereas only individuals who actually commit the relevant actions will pay the cost. This creates seemingly insuperable "free rider" problems that seem to make evolutionary altruism a terribly fragile thing.

The situation changes somewhat if different *groups*, rather than different individuals, are compared. For now individuals that commit the altruistic acts that favour the group can potentially be more fit than individuals who are in groups where no such acts are being committed. The "free-rider" problem, however, still remains. Groups where altruistic acts are being committed are always vulnerable to "subversion from within". "Free riders" in altruistic groups are fitter than altruists in the altruistic group, and hence evolution will, inevitably, favour the "free riders".

---

problem, see the section on "Units of Selection" in Ruse and Hull's (eds.) collection of readings *The Philosophy of Biology* (1998: 147-220)

3.6 Group Selection, Evolutionary Altruism and Simpson's Paradox

The above logic can, given certain conditions, be overcome. In order to understand how this is possible it is necessary to understand an apparent statistical paradox known as Simpson's paradox.

Consider the following scenario. A group of people, composed of a hundred men and a hundred women, are separated into two groups. The groups each contain at least one woman and one man, and each group consists of one hundred people. Assume that, as is often the case, the men are, on average, taller than the women. Simpson's paradox relates to an odd possibility that can arise with regard to the average height of the members of the different groups. It is possible that, despite the hundred men being taller than the hundred women, the average woman may still be taller than the average man within each group, considered separately. In other words, women are taller than men within each group, but men are taller than women if we average across groups. In this case what is true of the particular groups need not be true of the whole.

The above might appear to be impossible at first glance, but can be rendered intuitive by a simple example. Imagine one hundred men and women that vary according to height as one would commonly expect, so that the men are, on average, taller than the women. Now separate them into two groups. Let the one group be composed of the three tallest women and the ninety-seven tallest men. Let the other group be composed of the shortest ninety-seven women and the three shortest men. If the group we are considering roughly have the height-distribution that one would commonly expect, Simpson's paradox would probably arise. The two same-sex groups of ninety-seven will only marginally differ from the average of their sex. But the three shortest and longest probably deviate significantly from the average; there will be very tall women and very short men. Hence the three tallest women will probably be, on average, taller than the average of the ninety-seven tallest men. And the three shortest men will probably, on average, be shorter than the ninety-seven shortest women. In other words women will be taller, on average, within groups, while men will be taller, on average, if we average between groups.

It is important to note that Simpson's paradox depends on the distribution into groups resulting in a strong correlation between a characteristic of the group and the characteristics of the members of the group. The correlation in the above example was between group height and individual height; tall women were in the tall group and short men were in the short group[132].

This decoupling between the characteristics of the specific groups and the characteristics of the group as a whole can be used to explain how it is possible for evolutionary altruism to evolve. The following example (adapted from Sober (1998: 772-473)) shows how this is possible.

Consider two groups, both with one hundred individuals. Let the one group ("S – group") consist of one altruistic individual (an "A") and 99 selfish individuals ("S's"). Let the other group ("A – group") consist of 99 A's and one S. Remember that an altruistic trait is one that decreases the fitness of an individual relative to the group, but also increases the fitness of another individual in the group. Take the benefit from altruistic action to be a uniform function of the number of individuals performing the altruistic act. Imagine that, initially, an S in the A-group can leave 4 offspring, an A in the A-group 3 offspring, an S in the S-group 2 offspring and A in the S–group 1 offspring in the next generation.

The above figures were chosen so that an instance of Simpson's paradox is obtained. In both the A-group and S-group an S will be more fit than an A, yet A's are fitter overall than S's. Hence the trait of "being an A" will be selected for in the next generation. This arises because of the strong correlation between the type of individual and the group that the individual is likely to belong to.

The above, while helpful, does not tell the full story of how group selection can evolve. For, in both groups, the relative fitness of A's is declining. Each succeeding generation will also see a decrease in the absolute fitness of A's, until this becomes

[132] An interesting instance of Simpson's paradox occurred at Berkeley in the 1970's. The university was suspected of discrimination against women in its graduate admissions, because the percentage of women admitted was less than the percentage of men. However, a closer look at the facts showed that, in any given department, the percentage of men and women who were admitted was roughly similar.

negative. As evolution proceeds the A's will become extinct and S's will survive. Hence, ultimately, individual selection will triumph over group selection.

The above problem can, however, be counteracted if there is some assortative mechanism that counteracts the above process. If the groups are resorted after each generation so that like has a high enough chance of living with like, then this can counteract the evolutionary force operating at the level of individual selection[133]. The strength of group selection will then be a function of the efficiency of the assortative mechanism. If this assortative mechanism is foolproof , i.e. if all A's group together and all S's group together, then the *only* evolutionary pressure with regards to the S/A trait will be group selection. In other words, the A's will evolve and the S's go extinct. If the efficiency of the assortative mechanism is beneath a certain threshold then individual selection will outweigh group-selection and the A's will go extinct.

In the above model one can expect the usual evolutionary "arms race" to occur. Great ability at spotting A's or S's can be expected to evolve. This leads to an evolutionary pressure in favour of being able to cheat by pretending to be an A, while actually being an S. This leads to an evolutionary force in favour of an increased ability to spot cheaters, etc[134].

It has now been shown that evolutionary altruism can, because of between group selection, evolve under certain specific conditions. The ability of group selection to explain human preferences rests on the existence and efficiency of assortative mechanisms that have existed in our evolutionary past. Sober and Wilson (1998), whose formulation of group selection was used above, are scrupulously circumspect about making inflated claims for the importance of their model of group selection.

The apparent paradox was resolved when it was found that women tended to apply to departments with low acceptance rates, thereby giving rise to Simpson's paradox (Sober and Wilson, 1998: 25).

[133] The same effect can also be achieved if all groups fragment randomly once they have achieved a certain size. These fragmentations then need to happen often enough, and groups composed of selfish members only must be sufficiently unfit (*qua* group) so that they eventually go extinct. Provided the numbers are right, altruism will be selected for in such a case. See Sober (1998: 474).

[134] It is within such a context that an ability to deceive oneself, if it leads to an increased ability to deceive others, will be selected for, as was explained in chapter 5. It is also within such a context that blushing, as a signal of an inability to deceive, and, hence, as a signal of being trustworthy, will be selected for. See Joyce (2001: 144-145).

Yet they conclude as follows:

> The factors that can make group selection a strong force in the absence of genealogical relatedness seem to be abundantly present, especially in the small face-to-face societies that existed for most of our evolutionary history. Human groups do not invariably function as adaptive units... [nevertheless] most traditional human societies appear designed to suppress within-group processes that are dysfunctional for the group, and as a result natural selection has operated and adaptations have accumulated at the group level (1998: 192).
>
> At the behavioural level, it is likely that much of what people have evolved to do is *for the benefit of the group* (1998: 194).

It is beyond the scope of this thesis to present an argument for the correctness of the conclusions cited above. It seems clear that group selection, as defined, *can* occur. The historical *frequency* of its occurrence, however, is difficult to determine. It is part and parcel of an evolutionary process that it will tend to destroy evidence of its history as subsequent adaptations outreproduce previously advantageous adaptations. For this, and other reasons, Sober and Wilson make it clear that the conclusions cited above are based on partial evidence, and are not beyond correction by subsequent research. It does, however, seem to be a fair interpretation of the currently available evidence. And, for this reason, it must have a large impact on the current justifiability of any claims as to the type of preferences people are likely to have.

3.7 Group Selection as Cause of Psychological Altruism

The existence of evolutionary altruism does not necessarily imply anything with regards to the existence of altruistic preferences[135]. A being could, in principle, gain all the reproductive advantages to be had from group selection by having a desire to act so as to maximise the absolute "between-group" fitness of his genes. This would,

[135] The positive implications of group selection *vis-à-vis* altruism should not be overstated. It also has a dark side. While it will favour intra-group "niceness", it can also favour between group "nastiness" (Sober and Wilson, 1998: 9).

given sufficient computational capacity, have been the optimal design for evolution to hit on. In chapter 4 it was explained, however, why evolution cannot be expected to produce optimal design except by accident. Evolution is a boundedly rational designer that can be expected to be inelegant and intricate, but effective.

Group selection does help to build the case for the existence and prevalence of altruistic preferences because it is easy to see how a group selectionist agenda can be partially implemented by such preferences. Any altruistic preference that is to the benefit of the group, but not to the evolutionary benefit of the individual, could, in principle, be an instance of group selection. This could take the form of an all-purpose concern for the group of which one is a part, or a specific concern for some of the individuals in the group, or a distaste for harming the interests of others within the group through specific actions like stealing, lying, etc[136].

These preferences would then need to be complemented with preferences that allow the assortative mechanism needed for group-selection to operate. Here other-directed desires for the welfare of other altruists would be one mechanism whereby evolution can partly get the job done. The concern for the welfare of others will, presumably, result in interaction with them that is to their benefit. And, since "group" is defined in evolutionary biology in terms of fitness-altering interaction, and not in terms of some other criterion like spatial proximity (Sober and Wilson, 1998: 93), this interaction already helps to make me part of the other altruist's group[137].

Hence group selection gives reason to believe that evolutionary altruism might well be a historical reality. This provides support for the idea that a considerable number of altruistic preferences exist. And, since altruistic preferences can be reinterpreted as causes of pleasure, this implies that apparently altruistic preferences that have evolved *via* this mechanism cannot be used to object to the idea that people are egoist hedonists.

---

[136] A general concern with, for instance, honesty, might not count as psychological altruism, as the phrase was designed above, since it need not be an other-directed propositional attitude. Yet it might still have evolved through group selection. This demonstrates something important, namely that "altruism", as defined, is only one type of "niceness" that can be explained in terms of the processes outlined in this chapter.

[137] In order for the assortative mechanism to operate one could also have some decidedly un-altruistic preferences with regards to irredeemably selfish individuals.

The next argument in favour of believing people to have altruistic preferences derives from the way in which human interaction is structured. These structures are of such a type that they result in an evolutionary pressure in favour of other-directed, altruistic desires.

3.8 The Structure of Interaction as Cause of Psychological Altruism

3.8.1 The Iterated Prisoner's Dilemma and Tit-for-tat

The term "structure" will be used to refer to the relation between the different possible outcomes, calculated in terms of fitness, that an interaction, or series of interactions, can have for the parties involved in an interaction. The structures of some social interactions can result in outcomes that can be highly counter-intuitive. One of these counter-intuitive outcomes is that they can result in situations where "nice", or partially "nice" action can be surprisingly effective. This method of accounting for other-directed altruistic desires will be explained with reference to the most famous structure of interaction of them all – the prisoner's dilemma.

"Prisoner's dilemma" is the name commonly used in game theory to refer to a certain structure of social interaction, or "game", that has the following characteristics. Two participants have to make a choice without information regarding what the other party is going to do. Both parties can either "co-operate" with the other or "defect". The pay-offs of a prisoner's dilemma is its main defining characteristic. The possible pay-offs are of such a nature that the maximum gain for each party lies in the possibility of defecting, while the other party cooperates. Conversely, the worst possible outcome is to be a cooperator if the other party is defecting. For both parties it is better to cooperate if both are doing so, than for both to defect. What makes the prisoner's dilemma an interesting game is that, for each party, and no matter what the other party does, it is better to defect than to cooperate. However, both parties know this, and if they are rational, they will choose to defect. This, however, is a sub-optimal outcome as it would have been better for both if both parties had cooperated.

The peculiarity of the prisoner's dilemma is that the rational outcome, i.e. to defect, leads to sub-optimal outcomes. It is easy to construct examples of it. Consider, for

example, two people who decide to trade items the next day by, for example, leaving them at different places at the same time. Assume that there will be gains from trade if both parties actually keep their promise to leave the traded article. Assume both have to decide whether to trust the other without knowing what the other has done. Now, strictly in terms of material self-interest, it is apparent that, *no matter what the other does*, both parties are better off if they break their promise. But both sides are likely to realise this, break the promise, and the gains from trade are lost.

A similar situation occurs with the so-called "mortarmen's dilemma" (Ullman-Margalit, 1977: 30). It concerns the situation of two mortarmen in war who are under enemy attack. If both stay at their post when the enemy attacks they have a fair chance of warding off the attack and surviving. If both desert their posts then nothing stops the enemy from breaking through, and both have a low chance of surviving. But if one of them deserts, and the other stays to fight, then the deserter has an even better chance of living than if both had stayed to fight. In other words each of the mortarmen are better off deserting as long as the other stays to fight, but if both desert then they are each worse off than if each had stayed to fight[138].

The prisoner's dilemma was traditionally thought to teach us a very bleak lesson. Under some circumstances, the structure of interaction is of such a type that individual rationality leads to social loss. This changed greatly when Robert Axelrod, a political scientist, did a study concerning the optimal way to behave when two individuals play a series of prisoner's dilemmas against one another.

Axelrod's *The Evolution of Cooperation*, which included a chapter on the applicability of his ideas to biology by WD Hamilton of "kin-selection" fame, sparked an explosion of interest in the study of the iterated prisoner's dilemma as providing a basic model of some social interaction[139]. Axelrod's study showed that cooperation could evolve between self-interested parties, even if they are antagonistic.

---

[138] This type of problem is not uncommon in war. A technical solution mentioned by Ulmann-Margalit is the German practice in World War I of chaining soldiers to their guns. What is interesting is that soldiers would volunteer to be chained as long as their fellow-soldiers are also chained, thus assuring solidarity in battle (1977: 32).

Axelrod invited academics whose work was related to prisoner's dilemmas, to submit entries for a computer tournament where these entries would repeatedly "play" against each other in a round-robin tournament. These games were structured to have the same pay-off scheme as the standard prisoner's dilemma[140]. The counter-intuitive result of this game was that "tit-for-tat", an entry submitted by Professor Anatol Rapoport of the University of Toronto, was the clear winner. What was counter-intuitive about this is the fact that "tit-for-tat" does not attempt to exploit any other player. It is also laughably simple in comparison to the other entries.

"Tit-for-tat" begins each match by cooperating, and thereafter simply mimics the actions of the strategy it is playing against. If the other strategy is "nice" and keeps cooperating, then "tit-for-tat" cooperates in a like manner, and both strategies score highly. But if the other strategy is not "nice", then to such a strategy "tit-for-tat" is also not "nice", and both strategies get a low score. "Tit-for-tat" can never "win" in any individual match, it is designed to either get a draw or to lose by a slight margin. And yet it proved a decisive winner at the end of the round-robin tournament.

This was because of the inherent limitations of strategies that are "nice" and not "nice". A "nice" strategy will fare well against other "nice" strategies, but fare very badly against a strategy that is not "nice". A strategy that is not "nice" will do well against "nice" strategies, but poorly against strategies that are also not "nice". Whereas "tit-for-tat" is above all consistent, it either does well or loses marginally, this consistency allows it to win at the end of the day.

Axelrod repeated this tournament by inviting entries for a second round. All those invited to enter were given full details of the results of the first round. And here, surprisingly, the winner was again Rapoport, who simply resubmitted "tit-for-tat". Axelrod also held a tournament in which strategies played several rounds, but where each strategy's score in a preceding round determined how many copies of it ("offspring") would be in the next round, hence determining which strategies will be

---

[139] Thirteen years after publication this interest was still strong enough for Axelrod to declare the iterated prisoner's dilemma to be the "*E. coli* of the social sciences" (Axelrod, 1997: xi).

[140] It is important to note that these games did not specify a specific number of games to be played. In such a case reverse induction problems arise. Rather strategies "knew" there was a high probability that they would meet again.

dominant if passing through an "evolutionary" process. Here "tit-for-tat" again proved remarkably successful, "[b]y the one-thousandth generation it was the most successful rule and still growing at a faster rate than any other rule" (1984: 53).

How can the success of "tit-for-tat" be accounted for? Axelrod highlights four main features of successful iterated prisoner's dilemma strategies. They are "nice", meaning that they start by cooperating. They are "retaliatory", meaning they are willing to retaliate if they get exploited. They are "forgiving", meaning they will re-establish cooperation after mutual defections if the other is willing to cooperate. And they are "clear"[141], meaning that it is easy to identify what they are doing.

Axelrod uses these results to make certain recommendations to persons involved in a situation that has the structure of a prisoner's dilemma (1984: 110-123), and to recommend certain ways of transforming the setting of situations so as to encourage cooperation (1984: 124-141). These aren't of fundamental concern to the present inquiry. What is important to note is the following: cooperation can occur between antagonistic parties as a function of the durability and structure of interactions, if each party has a stake in the reactions of the other. Simply put, parties that don't like and don't trust each other might cooperate if they have a clear conception of their own interest.

3.8.2 The Iterated Prisoner's Dilemma, Tit-for-tat, and Psychological Altruism

If the prisoner's dilemma is considered from the perspective of evolution, so that the structure is defined in terms of evolutionary fitness, then a surprising fact emerges. The discontinuity between the actor as maximiser of evolutionary fitness and the actor as maximiser of preferences actually enables certain optimal outcomes to be achieved that would be otherwise impossible. A being that is in a prisoner's dilemma *qua* evolutionary actor need not be in a prisoner's dilemma *qua* fitness maximiser.

---

[141] Axelrod's use of "anthropomorphic" terms like "nice", "forgiving", etc. is criticised by FA Beer (1986).

Consider a being that has to decide whether to cooperate in a series of encounters, and imagine the evolutionary pay-offs to result in a typical prisoner's dilemma[142]. Imagine this being to have the psychological traits that Axelrod identifies as the main reasons for tit-for-tat's success. In other words it is "nice", and this "niceness" is an altruistic desire to be similarly nice to cooperative others. It is also "forgiving", and this character trait takes the form of an altruistic desire to reward those who have made amends for past "defection". Also assume that it is easy to observe their actions ("clearness") and that they have an aversion to defectors that causes similar defections ("retaliatory"). If the altruistic desires that could result in being "nice" and "forgiving" are built into their preference structure, then they are not, *qua*, preference maximisers, in a prisoner's dilemma at all. Rather they are maximising their preferences by being cooperative. As long as this preference for cooperative behaviour is safeguarded by an aversion to defectors that causes defection in kind, they can be viewed, in terms of evolution, as playing tit-for-tat.

Axelrod showed that tit-for-tat can be a quite robust strategy in iterated prisoner's dilemmas, and hence that it can evolve. This implies that the type of altruistic preferences that were mentioned above with regards to being nice and forgiving can be favoured by selection. Hence the structure of interaction between evolutionary actors can be of such a type that there will be an evolutionary pressure that can result in the selection of altruistic preferences. And, once again, if preferences are seen as the causes of egoist pleasure, this means that certain "altruistic" action can be seen as instances of egoist hedonism.

---

[142] These interactions do not necessarily have to be with the same being. In a population where everyone interacts with everyone else just once, and where the structure of these interactions yield a prisoner's dilemma, cooperation can still involve. Sigmund and Nowak showed that this is possible, given certain conditions, if the strategies have some knowledge of others' past behaviour and can then assign them an "image-score" (Badcock, 2000: 100-102). (This type of scenario also suggests how a desire to talk about past defections and cooperations, i.e. "gossiping" as something inherently rewarding, can be favoured by selection.) Boyd has also suggested that cooperation, where interaction between any two individuals is infrequent, is possible in an amended game where it is possible to punish both defectors and those who fail to punish defectors (Ridley 1996: 80-82). Axelrod (1997: 44-68) invokes similar "meta-norms" to explain how cooperation can evolve in *n*-person prisoner's dilemmas .

3.8.3 The Limitations of Tit-for-tat

The importance of tit-for-tat should not be exaggerated. It is used in this thesis only to demonstrate how the structure of human interactions can generate evolutionary pressures in favour of altruistic preferences. It is not the only structure of interaction that can have this effect. It is also not clear how much of human interaction can be said to have this type of structure. The reference to tit-for-tat is being used only as an example to demonstrate the theoretical possibility that the structure of human interaction can, in principle, give rise to strong evolutionary pressures in favour of the selection of altruistic desires, not that it will definitely do so[143].

It should also be noted that tit-for-tat, while remarkably robust, is not an unbeatable strategy in an iterated prisoner's dilemma of the type specified above. It's success, and the success of any strategy for that matter, is *always* to some degree a function of the others strategies that exist in the given environment. The situation also gets considerably more complex if some of the extreme simplifying assumptions regarding information in any environment are relaxed. For example, in games where strategies can make mistakes a clear problem emerges with tit-for-tat. Tit-for-tat will respond to a mistaken defection with a defection. If the competing strategy punishes defection, a string of mutual defections can result and tit-for-tat can be beaten by a generous strategy that forgives the occasional mistake without defecting in return.

One strategy that has been particularly successful in a more realistic game where mistakes can be made, tactics can be switched and all strategies are not defined in advance is a strategy called "Pavlov"[144]. Pavlov is also known as Win-stay/Lose-shift. It changes its behaviour if it gets one of the two worse pay-offs, but persists in what it is doing if it gets one of the better pay-offs. It is a "don't fix what is not broken" type of strategy. Pavlov is more vindictive than tit-for-tat in that it will continue to take advantage of a strategy that unconditionally cooperates, but is "nice" in that it will start by cooperating and forgive defections.

---

[143] The search for prisoner's dilemmas giving rise to the evolution of tit-for-tat strategies has been, at best, a partial success. There is strong evidence for tit-for-tat behaviour in vampire bats, and also in some dolphins, monkeys, apes (Ridley, 1996: 71) and viruses (Badcock, 2000: 91).

The success under different conditions of strategies like the more forgiving version of tit-for-tat and of Pavlov serves to strengthen the claim for the evolution of altruistic preferences. It is clear that the behaviour of the forgiving version of tit-for-tat can be caused by altruistic preferences. The nice behaviour of Pavlov can also be caused by altruistic preferences, even though these will have to be complemented by some less laudatory preferences in order to account for its exploitative streak.

3.8.4 A Note on the Relation Between Iterated Prisoner's Dilemmas and Group Selection

Group selection was discussed above as a distinct method of accounting for the evolution of altruistic preferences. Yet it is reasonably easy to see that the type of situation that causes group selection to operate can be seen as just another different type of social structure.

The above examples regarding iterated prisoner's dilemmas all concerned situations where the only competition is against agents that form part of one's own group. Group selection adds an important qualification to conclusions reached from such a study that, because of Simpson's paradox, it is not even always necessary to do better than the individuals in one's group in order to evolve.

One of the essential ingredients necessary for group selection to operate is that there must be some assortative mechanism that causes the different groups to be constituted in a highly nonrandom manner. This type of possibility can be incorporated in the iterated prisoner's dilemma by giving the strategies a choice about whether to interact with other strategies. Since the idea of a group is defined in terms of frequency of interaction, the option of choosing whether to interact amounts to an assortative mechanism that lets groups develop. Kitcher designed such a game and found that a nice strategy, discriminating altruism, achieves the same type of success that tit-for-tat achieves in the more standard game (Ridley, 1996: 81-82)[145].

---

[144] Pavlov's was originally designed by Anatol Rapaport of tit-for-tat fame, its strengths were first demonstrated by Sigmund and Nowak. For a discussion, see Ridley (1996: 76-80). On the limitations of tit-for-tat, also see Badcock (2000: 95-98).

[145] An intriguing question raised by such a game is the one regarding the ability of people to spot potential cooperators and defectors. Frank has found that strangers placed in a room for thirty minutes

In terms of this thesis any situation where the structure of interaction is of such a type that Simpson's paradox and an assortative mechanism is necessary for selection to occur is viewed as an instance of group selection. It should however be reiterated that, as was said in footnote 123, that I am not here making any substantial claims regarding problems concerning the proper "units of selection" or the relation of group selection to complementary ways of explaining the evolution of altruistic preferences. The categories chosen for the exposition of the main points of this chapter make little or no substantive difference to the points being made here.

3.9 Summary Concerning the Evolution of Altruistic Preferences

Three different processes whereby altruistic preferences can evolve have now been discussed. The first was kin-selection, which helps to account for altruistic desires toward genetically related individuals. The second was group selection, which helps to account for altruistic desires towards individuals that are, in some sense, seen as members of the same group. The third was the structure of social interaction. Here prisoner's dilemmas were used to demonstrate how the relative pay-offs of evolutionary interaction can give rise to evolutionary pressures in favour of the selection of individuals with altruistic desires.

The above explanations regarding the evolution of altruistic desires were intended to demonstrate that there is reason to believe that human beings have altruistic desires[146]. As such it helps to build the evidence for the main claim of this chapter. This claim is that our intuitive judgements as to what type of actions egoist hedonists would perform is untrustworthy. If a strong case for this contention can be made then the existence of "altruism" no longer serves as a strong objection to the idea that people should be egoist hedonists.

---

will do considerably better than chance at predicting whether any given person will cooperate in a one-shot prisoner's dilemma (Ridley, 1996: 82). Tooby and Cosmides (1992: 19-136) have marshalled considerable evidence in favour of the idea that the modular mind has a "cheater detection" module that enables it to be much better at seeing the logical implications of conditionals when these conditionals concern violations of the "social contract" than when they do not.

[146] The above reasoning is intended to explain *some* reasons why human beings can have altruistic preferences. It should not be taken to imply that all preferences are specifically determined by the genes. Consider, for example, a phenomenon like the conscience. The conscience causes a powerful affective response against certain action, and it will surely be uncontroversial to assert that the nature of these actions will partly be a matter of environmental factors.

The next way in which our intuitions regarding what egoist action would look like can be shown to be untrustworthy regards the social structure of the interaction of utility maximisers.

4. The "Structure of Human Interaction" Defense

Above it was explained why the structure of interaction between actors, viewed from the perspective of evolution, can give rise to evolutionary pressures in favour of the selection of altruistic desires. Evolution can solve this problem in a number of ways. It could, in principle, create a being that attempts to maximise fitness and for whom this is the only ultimate desire. Then it could provide this being with sufficient computational ability to figure out the structure of interaction and to realise that cooperative behaviour might well be optimal.

The other possibility would be to equip the being with a mixture of altruistic and other desires that result in the same behaviour. These desires, such as the desire to reward cooperators and punish defectors imply that the actor is not, *qua* utility maximiser, in a prisoner's dilemma situation. In the right context this method would have certain potential advantages. Simply desiring to reward cooperators and punish defectors is computationally cheaper than the whole process of analysing the structure of interaction and realising what possibility would be optimal. This would also be a potential pitfall, since a radical change in context could make the strategy a radically sub-optimal one.

If certain desires evolve that save on computational cost by making cooperative behaviour a straightforward matter of maximising preferences, then the converse of the above can result. Here it could now be possible for an actor to be in a prisoner's dilemma *qua* rational actor, but not *qua* evolutionary actor[147]. This provides a new

[147] Does the contrast between the actor as evolutionary agent, and as maximising agent, imply that the individual should be seen as "many selves"? The position defended thus far does, to a certain degree, take such a view, but does not treat such agents on a par. Rather, and consistently with the basic folk psychological categories used in this thesis, the conscious, maximising agent is a real agent. The agent as evolutionary actor is merely a set of processes that can be interpreted by using the intentional stance, an "as if"-agent. Ainslie (2001) gives an account of the "bargaining" that can occur between the many "selves" that constitute an individual. This can, in terms of this thesis, be seen as a fundamental conflict between the real, maximising agent, and the processes underlying maximisation that were selected for by evolution.

possibility of explaining how seemingly disinterested behaviour can be related to self-interest. For, in the exact same manner that the structure of interaction can give rise to situations where cooperative behaviour is surprisingly effective, so too cooperative strategies *qua* rational actor can have the same counter-intuitive effectiveness.

Everything regarding the surprising effectiveness of strategies like tit-for-tat to the actor as evolutionary agent also applies to the actor as preference maximiser. The difference, however, lies in that the game-theoretic structure of the interaction *qua* preference maximiser does not here give rise to altruistic desires. Rather it is possible that, if confronted with an action that appears altruistic, this apparent altruism is a mere consequence of egoism hiding behind the façade of a complex game-theoretical structure.

Note that, while this type of defense does presuppose the existence of a preference maximiser with a mind, it need not constitute an undue tax on the computational capacity of the actor. It is, off course, possible that an actor will figure out the structure of interaction and act on this knowledge. But it is also quite possible that the actor is simply imitating the actions of other social beings, conforming to the practice in a given culture, or using a method that worked in the past in a similar situation, etc. These practices can all have similar behavioural consequences, without the actor understanding much about the reason for the effectiveness of his actions. Behaviour that is similar to that of an actor having figured out the virtues of, for instance, tit-for-tat, can be the result of a set of beliefs that have behavioural consequences similar to desires spoken of above when considering the evolutionary actor. In other words the desire to reward cooperation and punish defection can often have the same consequences as the belief that one should cooperate and should punish those who defect.

These beliefs can still be traced to the structure of interaction if the reason for their adoption is based on their past success in precisely such a situation, and there exists some mechanism whereby the actor can learn from the past mistakes of herself or others, even without fully understanding the nature of such mistakes.

The next method of accounting for supposed altruism is related to the points made above, and concerns the cognitive constraints of human beings.

5. The Bounded Rationality Defense

It was already argued in chapter four that people can, due to the cognitive constraints imposed on them, the fact that decision-time is a cost, and the design-history of the mind, only ever be rule-followers as opposed to calculators. In other words, people will be satisficers and not perfectly rational actors.

This then has implications for how a specific act of altruism is to be interpreted. Even if it can be shown that a certain action, given the preferences of the actor and the structure of social interaction was sub-optimal, the bounded rationality of the actor still needs to be taken into account[148].

Consider a situation where cooperative action is mostly optimal, but with certain exceptions. This is the type of situation that was discussed in chapter four with reference to the hypothetical actors searching for gold coins. There it was argued that optimal action will often include unsuccessful action. This is because the cognitive constraints of the actor makes it too expensive to determine when to not commit an action that is usually unsuccessful. This implies that, even if it can be shown that to be a defector is sometimes optimal, it still does not necessarily follow that the actor can be adjudged an altruist for not defecting. Rather the simple rule of cooperating more than seems strictly optimal is rational, because of the savings on computational costs[149].

The above type of situation can result in excessive cooperation when judged against the standard of a rational actor. The converse of this is also possible. If defection is

---

[148] This could then give rise to an evolutionary pressure in favour of beings who can use such mistakes to their own advantage, This turn of events would give rise to beings who can spot such beings or can exclude such beings in some other way, etc. Here the usual evolutionary arms race would occur, as alluded to earlier in this chapter. It is important to realise that this type of evolutionary tinkering would not change the fundamental truth that we are satisficers. Rather we just become more and more complicated satisficers. Trivers' (1971) visionary essay alluded to earlier cites this type of arms race with regards to gaining advantage from relations of reciprocity and the resulting process as a possible cause for the sudden explosion in human cognitive development.

[149] This idea is also suggested by Wilson (1998: 486) and Ruse (1986).

often optimal, and the exceptions where one should cooperate are too computationally expensive to find, this will result in "surplus" defections. For this reason I wish to be careful about exactly about what is being claimed.

It is possible that the cognitive constraints of human beings and the nature of their environment causes more altruism than would otherwise be the case, if all interactions are considered. But it is also possible that, *if all cases are considered*, the opposite is true, i.e. that our relative stupidity results in more "selfish" behaviour than would otherwise be the case. But it does seem certain that there will be some *specific* cases where bounded rationality causes us to cooperate where we otherwise would not. This means that there exists the possibility that any given instance of supposed altruistic action can simply be the result of our constrained cognitive capacities.

6. The Cognitive Error Defense

The last way of accounting for apparently altruistic action is also the simplest. It is possible that the actor in question could simply have made some or other mistake. This mistake can be a simple one as to the nature of a situation or the nature of an actor. It could also be the type of simple error in reasoning that we all sometimes fall victim to. The negative consequences of such mistakes make it unlikely that an actor is likely to make systematic and persistent mistakes of this nature.

A more serious problem can arise because of mistakes with regards to ethics as such. If the central claim of this thesis is true, i.e. if people should be egoist hedonists, then the history of thought concerning the basis of morality is false. This means that the belief system of most people with regard to why actions should be committed is false. This is not an implication that only follows from egoist hedonist premises. The mere fact that there are a number of conflicting views with regards to the nature of morality guarantees, as a matter of logic, that most people who have had an explicit view on this issue have been wrong.

Such basic mistakes can wreak absolute havoc with regards to the actions resulting from them. Consider the case of someone being a rather crude utilitarian. There are certain actions that would be recommended by utilitarian premises that would conflict

with that recommended by egoist hedonism. And yet this would not constitute a refutation of egoism, rather these actions in question can be made consistent with egoism by showing them to rest on a factual mistake with regard to the nature of morality.

The implications of the above reasoning are not as disastrous as one would intuitively suppose. It is a curious fact, much commented on by writers in ethics, that the most divergent moral systems will often recommend the same type of action when applied to specific cases. The categorical imperative, utilitarianism, ethical intuitionism and virtue-ethics would all agree that a drowning stranger should definitely be saved if this can happen without any great danger to the actor. Some reasons for supposing that an egoist hedonist should also save the drowning stranger have also been advanced above. Hence the degree to which such fundamental mistakes, if indeed that is what they are, would cause sub-optimal action is less than one would expect. Yet, since all moral systems do not always agree, it does present a powerful way of explaining why certain apparently altruistic actions need not pose a challenge to egoist hedonism[150].

7. Conclusion

The main focus of this chapter was to argue that our idea of what egoist hedonist action would look like is untrustworthy. There are several ways in which action that appear altruistic can be shown to be consistent with viewing the actor as an egoist hedonist. It can firstly be shown that there would have been evolutionary pressure in

[150] The existence of altruism would seem to pose a challenge to my claims regarding egoism. In similar fashion, the existence of masochism seems to pose a challenge to my claims regarding hedonism. I think that problems regarding masochism can be dealt with in a way similar to how altruism was dealt with in this chapter. It should be noted that my main claim is with regards to ethical hedonism. No action can, in principle, falsify a normative doctrine, and hence no conception of masochist action can be a knockdown-argument against the normative thesis. This is not to deny that some account of why the normative injunction is not heeded is necessary. Here the following possibilities present themselves. Firstly, an act of apparent masochism can be instrumentally valuable with regards to obtaining pleasure. Secondly, the choice for the masochistic act can rest on a cognitive mistake, or be the result of the boundedly rational nature of the actor. A third possibility is more speculative. Ordinarily pain and pleasures are not the objects explicitly sought or avoided by the actor. Rather the actor will seek some other object, and the resultant experience will have a certain value as quality of the experience. It seems a possibility that, under certain circumstances, pain can enter into experience as the "object" of experience, without this necessarily affecting the value of experience adversely. This would be the converse of the common observation that pleasure often diminishes if one focuses one's concentration on the pleasure, instead of the object of the pleasurable experience.

favour of actors who gain pleasure from the welfare of others. This is because of kin-selection, group selection, and the structure of social interaction between evolutionary agents. The structure of interaction between preference maximisers can also, and for the same reason, often make apparently altruistic ways of acting surprisingly effective.

The other two ways whereby apparent altruism can be turned into egoism both involve our cognitive limitations. The first of these concerns our bounded rationality. There will be cases where simple cooperative strategies might well be optimal if our constrained computational capacities are taken into account. There also exists the possibility that a given action was based on a mistake. Since it is a simple matter of logic that most people throughout history have had mistaken beliefs regarding the nature of morality, this is a potentially very powerful way of explaining apparently troublesome actions.

The discussion in this chapter concludes the work in this thesis regarding specific rules and strategies that an egoist hedonist should follow. The next chapter will concern more abstract matters that have been implicit in the discussion thus far. These regard fact, value, and the nature of objectivity.

Chapter 7: Objectivity, Fact and Value

1. Introduction

The identification of value with the value of experience of a subject allows for the treatment of normative statements as a subset of descriptive statements. This view has certain epistemological implications; the first is that problems regarding the naturalistic fallacy are overcome. It now remains to demonstrate some of the further implications of such a view.

If a philosopher accepts the radical discontinuity between fact and value he seems to have three possible responses. The first is to treat this distinction as indicating that there are two different *types* of knowledge, both with their own logically distinct criteria of validation, i.e. "scientific" and "ethical". This thesis opposes such a view. Instead it tries to develop a conception of value that allows value-judgements to be treated as a subset of descriptive statements. This chapter will argue that this conception of the relation between fact and value allows for light to be shed on issues in philosophy concerning our understanding of "objectivity". It will be argued, simply put, that the conception of "objectivity" as "value-free" is a mistake. Rather "objectivity" should be seen as a value. It will also be argued that this does not have any impact whatsoever on the coherence and intelligibility of what is important about our traditional conception of objectivity. Simply put, even if objectivity is recognised to be a value, this does not, by itself, make it any less "objective".

The above conclusion can be shown to follow from a weaker assumption concerning the relation between fact and value than the one developed in this thesis. It does not only follow from viewing value-judgements as a subset of descriptive judgements, but also from any view that allows fact and value to be distinct. The idea they are distinct is both important to our understanding of objectivity and is fundamental to this thesis in that it is an assumption underlying the central claim being made. For this reason some of the most influential objections, mainly drawn from Putnam's *Reason, Truth and History,* will be argued to be baseless.

The chapter will end by considering one historically influential school of thought that seems to have sometimes supported its main conclusions by using arguments that rest on a misconstrual of the relation between fact and value. This school of thought is pragmatism.

2. The Relation Between Fact and Value

The claim that the "good" is a *quale* about which we can have true knowledge leads to a very simple view of the relation between fact and value. Value-judgements simply become a subset of descriptive judgements, capable of being assigned a truth-value in virtue of an entity that is independent of its representation in language. Determining the value of anything does not, if the above conceptualisation is accepted, raise any great conceptual difficulty. The value of a state of affairs is the difference in value, understood as a *quale*, between the possible world in which the state of affairs exists and the possible world that would result if the state of affairs does not exist. Any judgement as to this difference is true or false in virtue of its relation to the fact as to what the difference is.

3. Relevance of the Above View of Fact and Value to "Objectivity"

The notion of "objectivity", both as it applies to science and other activities, is a much contested idea in philosophy. A defense of this notion in its entirety is obviously not a matter that can be undertaken here. Yet there is one specific objection to the ideal of objectivity that can be shown to be groundless if the main idea of this thesis are accepted.

"Objectivity" is often, mostly within the realms of scientific practice, defined in terms of "value-neutrality" or "value-freedom". Here it is supposed that, when doing science, one should not bring one's prejudices to the laboratory. Rather we should just try and determine what nature is actually trying to tell us, without imposing our views upon her. It does not state that scientists should not make value-judgements, just that they should not make them *qua* scientists. Rather they should be reserved for the scientist's private life and the occasional letter to the editor.

This type of view of objectivity is not that difficult to upset. If objectivity is understood in terms of value-freedom, one just needs to show that the scientist has some goal that he is trying to achieve, hence he cannot be value-free. He cannot even object that he is just trying to ascertain "truth", for, as argued in chapter 5, the criteria whereby we determine whether to accept or reject a belief, i.e. truth, is itself based on a value-judgement. Even a seemingly "disinterested" search for truth still presupposes the scientist's judgement that "truth" is the type of thing we need more of.

The above reasoning seems to be valid, and will not be challenged. What will, however, be challenged is the view that some rather radical conclusions follow from these premises. For instance, it has been suggested that, since science cannot be value-free, "corrective biases" and "progressive political values" need to be introduced into science[151].

It will be argued that seeing "value-neutrality" to be impossible does not establish anything interesting. Rather the idea of "objectivity" as "value-freedom" was a bad conceptualisation of what is important about the idea of "objectivity". If one accepts that value-judgements are a subset of descriptive judgements, then realising the above conceptualisation to be false implies nothing with regards to debates concerning "relativism vs. foundationalism", or anything of the sort. Rather the traditional conceptualisation is interesting only as an example of how muddled thinking can be used in support of questionable ideas.

The above conclusion, however, does not only follow from viewing the good as a *quale*. Nor does it only follow from believing that there are representation-independent entities in virtue of which value-judgements are true or false. An improved understanding of "objectivity", one that safeguards what is important in the traditional notion from attack by the above argument, can be reached if fact and value are understood to be distinct.

The idea that fact and value can at all be distinguished has not escaped criticism. If they are not conceptually distinct then the implications for this thesis are dire. The

[151] This characterisation of "social constructivist"-doctrine is found, and excellently criticised, in Koertge (1998: 4).

idea that values can be a subset of facts clearly rests on the assumption that some distinction can be drawn between fact and value. Hence this issue will be examined below in some detail; both because the distinction between fact and value is vital to this thesis and because it allows the notion of "objectivity" to be clarified.

4. The Distinction Between Fact and Value

4.1 Definitions

At the start of this thesis, "fact" was defined as the extra-linguistic reality, in virtue of which a proposition's truth-value is determined. Note that this definition does not commit one to any particular definition of "truth". If one defines "truth" as *correspondence*, then the truth-value of a proposition is determined by the corresponding fact. If one defines "truth" as *coherence*, then "facts" constitute the ontological commitment demanded by a given theory if the theory is to be consistent. "Value" was be defined as the irreducible[152] goodness or badness of an object, action, etc. A "value-judgement" is then a judgement regarding the irreducible "goodness", "badness", etc., of something, or the equivalent judgement that something "should" or "should not" be done.

Using these definitions, a "standard account" of the case for the seperability of fact from value can be given. States of affairs can be asserted to have certain properties. One of these properties is the goodness or badness (however understood) of the state of affairs. In this manner I can describe a certain act of killing as morally wrong, as having been committed by Jones and as having been done by using a knife. I can go further by saying that this information is valuable to the police, that it is bad for Jones if the police have this information, etc.

Here factual judgements and value-judgements can apparently be clearly separated. In this example the fact that Jones committed the act of killing stands in no logical relation to the assertion that this was done using a knife, i.e. there is no apparent

[152] "Irreducible" is here used to exclude any use of "good" or "right" that can be defined without using "value-terms". Hence cases where something is "good" or "bad" inasmuch as it satisfies a given goal, or the logically equivalent "conditional/hypothetical should", are excluded.

contradiction involved in agreeing that Jones did commit the act, but denying that it was done using a knife. In a similar manner the assertion that Jones committed the act of killing stands in no logical relation to the assertion that this was morally wrong, i.e. there is no apparent contradiction involved in agreeing that Jones did commit the act, but denying that it was morally wrong. In a similar manner it is logically possible that Jones did commit the act, but that this information is useless to the police, etc.

I wish to make it clear that it is the fact/value *distinction* that is being defended here, not the fact/value *dichotomy*. In other words, I am defending the idea that facts and values are ontologically distinct, which implies that factual judgements can be separated from value-judgements. The idea of a fact/value dichotomy seems most often used to refer to the idea that only factual judgements refer, whereas value-judgements are non-cognitive. The fact/value dichotomy depends on the intelligibility of the fact/value distinction, but asserts something extra. This is that the objects of a value-judgement (i.e. irreducible goodness or badness) are not instantiated. This thesis tries to present an argument that allows the distinction to be maintained, but escapes this dichotomy.

Hence the cognitive status of value is not being pre-judged. If value-judgements are descriptive, i.e. there is a matter of fact that determines the truth or falsity of value-judgements, then an assertion of value refers to a fact about a given object. Facts about values are then - to use an example from Pigden (1993: 421-431) - a perfectly delineable *subspecies* of facts in general in the same sense that facts about hedgehogs are a perfectly delineable subset of facts in general. No one has ever doubted that the fact/hedgehog distinction is dubious. If naturalism or intuitionism is true then the fact/value distinction is of the same type.

If value-judgements are non-descriptive, i.e. there is no matter of fact that determines the truth or falsity of value-judgements, then the situation is even clearer. In the case of a fact/value dichotomy, the realm of statements that can be true or false in virtue of a fact necessarily excludes value-judgements. No clearer case for the strict separation of fact from value can be asked for. Whether value-judgements are then taken to be nonsense, emotive utterances, commands, prescriptions, intersubjective rules of

action, etc. does not matter. The idea of descriptivity provides an insuperable basis for the fact/value distinction.

If the case for the separation of fact from value is as strong as it was portrayed to be above, then how is it possible that some very capable philosophers can ask us to drop this distinction? There are three possibilities that either suggest themselves or occur in the literature. These will be examined in turn.

4.2 "Entanglement 1": "Truth" is Identical to "Value"

The separation of fact from value portrayed above might be seen as circular. "Fact" was defined in terms of extralinguistic reality and truth, whereas value was defined as a property of facts. It might be argued that the above definitions beg the most important question, namely whether it is possible to form a conception of facts or truth independently of value.

A pragmatist might well make this type of objection. William James stated that the concept of truth is "essentially bound up with the way in which one moment in our experience may lead us towards other moments which it will be *worth while* to have been led to" (1978: 98; my italics). Here truth is explicitly tied to the idea of being valuable. Indeed James makes the famous pronouncement that there is no difference between saying "it is true because it is useful" and "it is useful because it is true" (1978: 98).

If "useful" and "true" are synonyms[153], then separating fact and value would be like separating "bachelors" from "unmarried men". Here fact and value are not "entangled", nor is the distinction "blurry", but there simply are no two things to become entangled. Rather statements have a property, the "degree to which they are useful to believe", and saying that a statement is true or that there exists a fact in virtue of which it is true is just an inelegant and/or misguided way of again affirming the existence of this property. Here the fact/value distinction disappears, not because

[153] Rorty sometimes seems to endorse this view of "truth" as merely meaning "useful": "'P' and 'we are better off even now if we believe 'p' come pretty close, for pragmatists, to saying the same thing." (Rorty, 1997: 19).

facts and value are inseparable, or because of some subtlety regarding value-judgements, but because there simply are no facts in the sense defined above.

This is a clear instance of the fact/value distinction being "overcome", but only for the rather brutal reason that facts are thrown out the window. Such a crude version of pragmatism removes all constraints from the process of belief-formation. If it has any validity then it might seem good news indeed, for now everything that I wish to be the case is, *by definition*, true. But clearly people hold any number of beliefs that they would have preferred to be otherwise. Hence there must be at least one constraint on belief-formation over and above "usefulness". Regardless of what this constraint is taken to be, it can be used to give content to the notion of "facthood".

It is doubtful whether any philosopher has ever defended, or consistently defended, the crude pragmatism outlined above. Rather James is sometimes interpreted as adding constraints like "correspondence" and "coherence" [154], while Rorty would also add the constraint of being able to justify my beliefs to other people (Rorty, 1997: 9). Any argument regarding the validity of such positions fall outside the scope of this thesis. But note that, if one believes "truth" to name a set of constraints on belief-formation, there is again no reason to believe that facts and values are, in any sense, hard to separate.

Holding "truth" to be a set of constraints on belief-formation amounts to saying that beliefs can "correspond", "cohere", "be valuable" and "be generally agreed upon", and that "truth" is simply a name for one or more of these properties. Statements that affirm that a given belief is valuable, while denying one or more of the other properties, are evidently not tautological or self-contradictory. Hence there does not appear to be any *prima facie* reason to think that they are not conceptually distinct, and to abandon the "standard account" of the fact/value distinction as outlined above. There would be a problem about the tautological or self-contradictory nature of such statements *only* if one of these properties was simply a confused synonym for "value". This is the position that was sketched above, and, if anyone seriously holds it, they

[154] This interpretation of James is criticised below.

might rightfully object to the "fact/value-distinction". But it is probably fair to say that the *onus* of defending such a position would rest on them.

If pragmatists – and others – do not doubt the fact/value distinction based on an *identification* of "value" with "truth", then what does this doubt rest on?

4.3 "Entanglement" 2: Certain Factual Claims are Also Value-claims

4.3.1 Putnam's Argument

Putnam is another pragmatist who insists that the distinction between facts and values is "fuzzy"[155]. He sets forth two arguments that will be considered here, the first rests on the existence of expressions like "He is inconsiderate" or "She will do anything for money". These expressions clearly convey some information about someone, but it is also fairly clear that they convey a negative valuation of someone.

Before considering his argument, it should be noted that this phenomenon does not, *of itself*, present any reason for abandoning the "standard account" set out above. If fact and values are believed to be distinct, then this is no reason to be surprised at the existence of terms that can be used to assert both something about value and fact. There are many terms that can be defined by using conceptually distinct elements. In this way "bachelor" can be defined as "unmarried" and "man", "lawyer" as "practitioner" and "law", etc. There is no reason to suppose that there cannot, similarly, be terms that are defined in terms of "good/bad" and of the "something" being called "good"[156]. Hence the easiest definition of cases like "inconsiderate" would be to distinguish between "good" and "the actions that are being called "good". Or to explain "She will do anything for money" in terms of "the actions she would commit" and "badness". Or, to put it more generally, to treat the contribution of any such term to a sentence in terms of its "object" and the "evaluation of this object".

---

[155] His arguments, the clearest for such a contention that I have found, are set forth in Chapter 6 and Chapter 9 of *Reason, Truth and History* (1981).

[156] Instances less innocuous than "inconsiderate" – in the sense that they can smuggle normative claims into descriptive arguments – would include "normal", "natural", "reasonable", "traditional" and the like.

The above analysis is one Putnam labels the "two-components theory" (1981: 203), and rejects (203-205). His reasoning is as follows: In order to separate the "evaluative" from the "factual", one would need to restate the "factual" in language that does not contain any "evaluative claims". However, any analysis into the language of physics (as supposedly "value-free" language *par excellence*) would fail to capture the meaning of "inconsiderate". Any "particles in motion" description of, for instance, "action" would not mean the same as "action"[157].

The same goes for any "...ordinary language predicate whose conditions of application do not mesh well with those which govern physical concepts. What this means is that, if there are two components to the meaning of 'X is considerate', then the only description we can give of the 'factual meaning' of the statement is that it is true if and only if X is *considerate*. And this trivialises the notion of a factual component" (205).

Putnam's objection is that, if we wish to separate the "factual" from the "evaluative" content of a term, we can only state the "factual component" by employing the term itself. Any such attempt would clearly have to assume that such a separation is possible, i.e. that fact and value are ontologically distinct. Hence it cannot be presented as an argument for such a separation.

4.3.2 Reply to Putnam's Argument

The first and obvious objection to Putnam's opposition to the fact/value distinction is that he is constantly separating "fact" from "value" himself[158]. In fact, the very ability to come up with examples that have both factual and normative components seems

---

[157] For reasons of brevity I have not used Putnam's example. The above is, however, intended to be equivalent to his reasoning.

[158] Putnam admits this *prima facie* objection, in that he repeatedly refers to "...two factors, rational acceptability and relevance" (202). But he then states that they are "interdependent in any real context...". What constitutes a "real context" is left unexplained, one might well ape the Austin of *Sense and Sensibilia* and ask what an "unreal", "fake" or "imaginary" context might be. The statement does not seem to mean anything more than an assertion that there are contexts in which words convey both fact and value, something no one would seriously dispute. (Putnam does add that the use of any word, etc. involves one in a "history, a tradition of observation, generalisation, practice and theory" (203). This, however, is an unrelated argument, discussed in (4.4) below.)

impossible to explain unless Putnam has a reasonably clear notion of "fact", "value", and the conditions under which they are both present.

The more serious objection is to his assertion that we can only specify the meaning of terms like "considerate" by using terms like "considerate", which is not at all helpful. But consider the statement "It is good to be inconsiderate". This might well, on the surface, appear to be a statement that is self-contradictory, since it ascribes "goodness" to something that is, by definition, "bad". Such a statement might well seem akin to saying that "Circles are square" or "bachelors are married".

But there is an important difference between "Circles are square" and "It is good to be inconsiderate". Someone who states that circles are square will merely be met with puzzled stares. Such a statement is unintelligible; all that can be deduced is that the person does not know the meaning of "circle" or "square". But the statement "It is good to be inconsiderate" does not need to be interpreted in like manner. In fact I would venture that very few people would be at all mystified by the *meaning* that the speaker is trying to convey.

For clearly the speaker can be interpreted as saying that certain actions, which we refer to as "inconsiderate", are good or should be committed. Here the speaker is simply *disagreeing* with the valuation of a certain set of actions. And, while one might wish to quibble about his use of the term "inconsiderate", this is not necessary for a perfectly intelligible argument about the moral worth of those actions we normally call "inconsiderate" to occur.

How does the above affect Putnam's argument? If, as Putnam says, the "evaluative content" of "inconsiderate" is inextricably tied to the "factual content", then a disagreement about the moral worth of "inconsiderateness" would be *unintelligible.* Simply put, the statement "inconsiderateness is bad" would be tautological and the statement "inconsiderateness is good" would be self-contradictory. And any argument about the "badness" of "inconsiderateness" could only be an argument about the rules governing the use of the term "inconsiderate". But, clearly this is not the only interpretation of the statement "inconsiderateness is bad". Using a weak hermeneutical principle of charity, one which only assumes the speaker to be a

reasonably competent user of language, the statement "inconsiderateness is bad" can be given a non-analytic interpretation. This non-analytic (not tautologous or self-contradictory) interpretation of "inconsiderateness is bad" is *only possible if the evaluative element can be separated from the non-evaluative component*, as is claimed by the "two components theory".

This presents the possibility of answering Putnam's challenge, without resorting to what he calls the myth that scientific description is the "One True Theory" (1981: xi). The "factual meaning" of "inconsiderate" can simply be defined as the meaning of the term "inconsiderate", *as used* in the phrase "inconsiderateness is good". Whether these conditions can also be stated in any other way is simply beside the point. The intelligibility of "inconsiderateness is good" *logically guarantees* the existence of non-evaluative uses of "inconsiderate". And these uses are necessarily distinct from evaluative uses, otherwise the intelligibility of "inconsiderateness is good" would disappear.

The same can be done with any term/phrase that has both normative and factual content. It is only possible to ask about the value of anything if the "goodness" or "badness" is conceptually distinct from the "thing, action, etc." being called good or bad. And in each case the non-evaluative meaning of the "thing, action, etc." can be defined as the conditions for its correct application in a non-analytic proposition affirming its "goodness" or "badness"[159].

Putnam does present a further argument for the contention that fact and value are always intertwined. This is in many ways a more interesting argument, and one that is regularly presented by authors who wish to deny the fact/value distinction.

---

[159] Putnam uses "inconsiderate" as a paradigm example of fact and value being mixed up, but seems to indicate that they are always mixed up, only more clearly in the paradigm examples (1981: 201-205). If the idea that they are always mixed up is taken seriously, *then all moral discourse becomes impossible and unintelligible.* For it implies that *any* statement about the goodness or badness of something can only be tautological or self-contradictory. Hence *any* dispute about the moral status of controversial practices, say abortion or cloning, can *only* be seen as an argument about the correct uses of the words

5.1 "Entanglement" 3: All Factual Claims Presuppose Value-claims

4.4.1. Factual Claims Imply Value-claims

At the start of this chapter it was claimed that, when descriptive claims are made, the maker of these claims is also necessarily committing himself to any number of value-claims. Putnam states this point as follows:

> Take the sentence 'the cat is on the mat'. If someone actually makes this judgement in a particular context, then he employs conceptual resources – the notions 'cat', 'on', and 'mat'- which are provided by a particular culture, and whose presence and ubiquity reveal something about the interests and values of that culture, and of almost every culture. We have the category 'cat' because we regard the division of the world into *animals* and *non-animals* as significant, and we are further interested in what species a given animal belongs to. It is *relevant* that there is a *cat* on the mat and not just a *thing*. We have the category 'mat' because we regard the division of inanimate things into *artifacts* and *non-artifacts* as significant, and we are further interested in the *purpose* and *nature* a particular artifact has. It is relevant that it is a *mat* that the cat is on and not just *something*. We have the category 'on' because we are interested in *spatial relations* (1981: 201-202).

It seems certain that behaviour can, in a certain sense, be said to reveal certain values, preferences, etc. of the actor. If we make value-judgements, then they must, in some way or other, be reflected in our actions. And, since "distinguishing", "speaking", "claiming", etc. are *species* of action, knowledge must in some sense reflect value-judgements.

"abortion" or "cloning". Surely such an unavoidable absurdity is a *reductio ad absurdum* of the idea that fact and value are always mixed up.

4.4.2 Relevance of the Above

While it certainly is, in some sense, correct to state that all factual claims presuppose a commitment to some normative claims, it is not immediately clear why this matters. Does it imply that fact and value are inseparable? If someone simply says that facts and values are "entangled", and by this *only* means that all factual claims presuppose normative claims, then that person is, as stated above, correct. This sense of "entangled" is, as will be shown below, epistemologically trivial.

But, if factual claims and value-claims are "entangled" in the above sense, there is also another sense in which they are distinct. This appears, at first glance, to be obvious. Putnam goes on at some length to outline the value-claims inherent in 'the cat is on the mat'. If these were not distinct the passage quoted above would be unintelligible.

More importantly, the above reasoning is often used to argue for a sense of "entangled" different from the above. This can be best seen by looking at Hans Albert, who, after stating that factual claims necessarily presuppose value-claims, draws the following conclusion:

> [I]n the last resort the unfoundable character of value statements must itself colour knowledge (1985: 79)

> It is therefore impossible to base a thesis of the irrationality of decisions upon their unfoundability without this thesis necessarily being extensible to the whole realm of knowledge (1985: 79).

Does the fact that all factual claims presuppose normative claims imply that any potential problem with the cognitive status of normative claims must necessarily result in a problem with the cognitive status of factual claims? The claim that this is the case seems to be the main reason why the "factual claims presuppose normative claims" – argument is regularly presented by authors who wish to "transcend" the fact/value distinction. In similar vein the fact that scientists make value-judgements

(choosing what to study, using distinctions, etc.) is commonly used to claim that there is an important sense in which scientific knowledge is not "value-neutral".

That all of knowledge presupposes value-judgements seems reasonably clear. But the idea these value-judgements necessarily "colour" our knowledge, or, that any doubt as to the validity or cognitive status of value-judgements has implications for the rest of knowledge, is false, as will be shown below.

4.4.3 Factual Claims "Presuppose" Normative Claims

The use of "presuppose" in the above heading is ambiguous. As was discussed above, any action, including making factual claims, implies that certain normative claims are tacitly being accepted, or govern my actions. However, this does not imply that these factual claims *logically depend* on these normative claims.

If I claim that 'Socrates is mortal', then I am logically committed to the claim that 'at least one thing is mortal'. The falsity of the latter would be inconsistent with the truth of the former, hence the former can be attacked by attacking the latter. This is one sense of "presuppose", but it is not the one used in the statement 'factual claims presuppose value-judgements'.

If I make the claim 'Socrates is mortal', then I am acting in accordance with some value-claim, imagine it to be 'It is important to assert that "Socrates is mortal"'. While the assertion of the former might commit me to the latter, and necessarily commits me to something akin to the latter, the truth of the one does not logically depend on the truth of the other. A situation can be imagined where Socrates is immortal, but it is still important to assert that he is mortal, i.e. to lie. Or a situation where Socrates is mortal, but I am mistaken in thinking that it is important to assert this. It could be the case, for instance, that I am having an argument with someone about Socrates' mortality, but that a fire has broken out and I would be best served by keeping quiet and running away. Hence, unlike 'Socrates is mortal' and 'At least one thing is mortal', the truth of the one statement is logically independent of the truth of the other.

Simply put, the value-judgements that a factual claim commits me to *are not supporting premises* of that factual claim. Hence even a successful attack on the value-judgements of a speaker still leaves the factual claims of the speaker untouched.

This is related to the basic insight behind the naturalistic fallacy, the realisation that, as a matter of logic, "is" can imply "is", and "ought" can imply "ought", but the one cannot *logically* entail the other. And it doesn't matter how many factual claims are made, or how many normative judgements they presuppose, for "never the twain shall meet". Hence I can say that "the cat is on the mat", and admit that I am committed to all the value-judgements Putnam found implied in such a statement. But, since these value-claims are not supporting premises of my factual claims, they can all be wrong, or nonsensical, etc., without this having any impact whatsoever on the truth of 'the cat is on the mat'. The lack of logical entailment between "is"-claims and "value"-claims provides an insuperable wall between these two types of statements.

4.4. "Supervenience", "Prejudice", etc.

Above it is argued that the "factual claims presuppose normative claims" – argument fails to provide any support for thinking that the nature or truth of value-judgements necessarily has implications for the nature or truth of factual judgements. I say "*necessarily* has implications", however, since the one does, all too often, determine the content of the other.

Note that this does not, necessarily, have to be an illegitimate move in an argument. If "moral qualities" are found to be natural qualities, or supervenient on natural qualities, as is argued in this thesis, then the existence of a natural quality can indicate the existence or non-existence of a natural quality. Here, while there is no logical relation of entailment, the relation of supervenience legitimates the inference. In fact, in an argument it would amount to the exact same thing as a "normative premise".

Does this present any problem to the fact/value distinction? At the start of this chapter it was stated that the fact/value distinction is intelligible *regardless* of what value is taken to be, this is also the case here. While an argument *might* be formulated that hinges on a relation of supervenience, it is not a necessary characteristic of knowledge

that it hinges on such relations. Whether the supporting premises presented for any given statement include the assertion of a relation of supervenience is a matter of fact. We can, and do, formulate arguments that do not rest on such assertions. Sensibly, since the idea of "supervenience relations" is somewhat mysterious, such appeals are normally outlawed in academic discussions. But, irrespective of whether they are legitimate or not, they *can* be coherently outlawed, since all conclusions do not necessarily have such an appeal as a supporting premise. Hence nothing about knowledge in general follows from the fact that, on some conceptions of value, an inference from fact to value and *vice versa* might be legitimate[160].

A more lamentable situation occurs when the rational reconstruction of someone's thought-processes includes the assertion of some relation, supervenient or otherwise, between fact and value. Consider something like: "It would be bad if man evolved from apes, hence it cannot be true", where the implicit premise is that "the truth about life cannot be bad". It would probably be fair to say that all people, at some point, fall foul of this type of reasoning, we commonly refer to it as being "prejudiced". But, while this type of reasoning is definitely possible, it is, as shown above, not necessary for an argument to contain such premises.

4.5 "Truth"

If more than one definition of "truth" is intelligible (which amounts to saying that there can be different constraints on belief-formation) then the specific criteria used by an individual or community represents a *choice*. In other words "truth" is a norm or value.

Does the idea that the constraints on belief-formation are chosen affect the fact-value distinction? The answer is no, and the reasons are the same as were offered above. If a choice for a specific criterion of "truth" presupposes an assertion of the value of this criterion, then this assertion is still not a supporting premise of any claim regarding the intelligibility of such a criterion, or any instance of its application. I might be

[160] Even if such a move is allowed, there is nothing particularly problematic about it with regards to the fact/value distinction as such. If evaluative qualities supervene on "natural qualities", then they are conceptually distinct *by definition*.

wrong about the *value* of "correspondence", "coherence", consensus", etc., but this is logically unrelated to whether beliefs in general can "correspond", "cohere", etc., or whether a specific belief "corresponds", "coheres", etc[161]. Simply put, there is no more reason to suppose that the "truth about truth" is entangled with the "value of truth" than there is to think that the "truth" of "Socrates is mortal" is entangled with the "value" of asserting that "Socrates is mortal"[162].

I wish to insert one clarification so as to be avoid confusion as to what is being claimed here. One major topic with regard to the distinction between fact and value was not discussed in this chapter. Our knowledge is, for the most part, fallible. While some philosophers have claimed that we can, regarding some types of propositions, have certain knowledge, no one would claim this for all our knowledge-claims. Rather our beliefs are based on evidence, and can be upset by new evidence. Here a problem arises, for clearly a set of rules needs to be employed to decide when evidence is sufficient to warrant holding a belief to be true. The status of these rules, i.e. whether they are "values" or only appear to be so, has been extensively debated[163]. This topic will not be discussed here.

This chapter concerns the relation between fact and value. The issue regarding the relation of evidence to belief is, in principle, a distinct issue. Here the question concerns the relation between "justified belief" (or "putative fact") and value, not fact and value. Simply put, the question "when is a belief true?" and the question "when am I justified in believing that a belief is true?" are distinct questions.

---

[161] The long association between pragmatism and "coherence" or "consensus" – theories of truth might be partially explained as resting on a misunderstanding of the above point. Unless pragmatism is defined as any doctrine that dismisses the idea of correspondence, there does not seem to be anything specifically *pragmatic* about viewing "coherence" "consensus", "simplicity", etc., as the proper criteria for belief-formation. The association between pragmatism and "coherence", etc., might well be due to the fact that these are "values". But if this criterion is used then "correspondence" is equally pragmatic, since it is no less a value than "coherence", "consensus", "simplicity", etc.

[162] If different sets of criteria for belief-formation can be chosen, then people can have contradictory "true beliefs", and this contradiction is then the result of evaluative judgements. This 'contradiction', however, is merely apparent or verbal, since the phrase "true belief" here has two different meanings.

[163] See, for example, the essays in Brodie (1970).

6. The Distinction Between Fact and Value and Objectivity

If the distinction between fact and value is allowed, then what is important about our ideal of objectivity can be stated without recourse to the idea of "value-freedom". At the start of this chapter it was explained that the definition of objectivity as value-freedom makes it easy to reject the notion. But, in light of what was said above, it must surely be clear how this problem can be overcome. Putnam's argument to the effect that all our knowledge includes value-judgements was conceded to be correct, but Albert's argument as to the implication of this argument was dismissed. The fact that we make any number of value-judgements when we use language and pursue knowledge does not render these value-judgements supporting premises of any of our knowledge-claims. The same goes for objectivity. If there is anything coherent about our ideal of objectivity, then the fact that objectivity is a value cannot be used to attack the ideal of objectivity. Any assertion regarding the coherence or intelligibility of objectivity is logically unrelated to any assertion as to the value of objectivity, if fact and value are distinct.

If someone merely *defines* "objectivity" to mean "value-free" then the argument cannot be challenged. But consider what we mean when we call a scientist "objective", and especially what we mean when we call someone "not objective". I think it is uncontroversial to say that this is commonly used to condemn someone for being "biased" or "prejudiced". And what is meant by this? It will surely be uncontroversial to say that we blame someone for not letting the object of his inquiry decide a certain matter. Instead the person lets his own interests shape his view or decision on a certain matter[164].

In other words the difference between being "objective" and being "value-neutral" can be viewed, not as a difference between making a value-judgement as to which belief to hold and not making a value-judgement, but as a dispute between two *different* value-judgements regarding the criteria involved in belief-formation. Even if

[164] Schlick is clearly referring to this when he states that: "Desire for the truth is the only appropriate inspiration for the philosopher when he philosophizes; otherwise his thoughts run the danger of being led astray by his feelings. His wishes, hopes and fears threaten to enroach upon that objectivity which is the necessary presupposition of all honest inquiry" (1962: 2).

the value of objectivity can be deduced from a commitment to truth, then the fact that "truth" is a value still negates the definition of objectivity as "value-free".

The following question arises based on the above reasoning: how can one reconcile the apparently contradictory claims that the subject is always driven by self-interest, and yet is able to admit facts he would have preferred to be otherwise? To a large degree this was already done in chapter 4 and chapter 5. There it was explained that people should follow rules, and it was argued that it is in a subject's self-interest to let "truth" be a norm of inquiry[165]. Hence it is in a subject's self-interest to let his expectations be corrected by reality. The discontinuity between an ultimate goal and the sub-goals whereby an ultimate goal is best pursued removes the apparent contradiction. Trying to ascertain "truth" is a self-interested act. The act of "letting reality correct my expectations" was mentioned above as what is important about our notion of "objectivity". Choosing to do so is to judge it likely to lead to successful action, hence it should be apparent that objectivity is *itself* a value.

"Objectivity" can loosely be defined as a habit of mind or willingness to admit evidence[166] that the subject would have preferred to be otherwise. This act of "admitting evidence that you would have preferred to be otherwise" is not "disinterested" or "value-free", but simply the best way to pursue the self-interest of the subject. Note that, in accordance with the above discussion regarding value and truth, the fact that "objectivity" is a value does not make it any less "objective". The value of objectivity, as was the case with truth, is distinct from its coherence and intelligibility.

Hence the situation is now reversed. Where value-judgements were seen as the enemy of objectivity, self-interest now becomes the *reason* for objectivity.

It is important not to overstate the case made here. I am not implying that scientists are always objective. As especially Kuhn has so eloquently demonstrated, objectivity is an ideal that is sometimes imperfectly followed. Neither is it being asserted that this

---

[165] Note that the argument here does not, strictly speaking, rest on the idea that "truth" should be a norm of inquiry, just that it is not logically incoherent to suppose that it is.
[166] This description means that "objectivity" is a *methodological* criterion.

totally clarifies the problem of objectivity. Opponents of the idea tend to claim they find it unintelligible. This is not an issue that will be discussed. What is being claimed can be summarised as follows: the interests and value-judgements of the observer can cause him to be biased, but it is also the basis of his "objectivity". Any argument that uses the interest of the observer to make a case against the very *possibility* of objectivity fails to understand how self-interest can lead to rule-following behaviour.

7. Objectivity, Truth and Pragmatism

In this thesis I have at various times paused in order to relate the analysis being made to various issues and movements in philosophy. I wish to do so again in order to relate the above idea to the pragmatism of James and Rorty.

Above it was argued that the idea that truth is valuable need not conflict with the idea that truth is objective. Rather the distinction between fact and value, added to the distinction between an ultimate goal and sub-goals, allows these two ideas to be consistent. An inability to see that these two ideas are consistent seem to have lead to more confusion than the issue merits.

I wish to clarify exactly what will be argued below. It will not be argued that pragmatism is false. An examination of pragmatism falls outside the scope of this thesis. What will be argued is that one of the ways whereby pragmatist conclusions are typically reached is fallacious, and that this way of reaching such conclusion seems to rest on an inability to reconcile the "value" of truth with the "truth" of truth. In other words, it is the same argument found in Putnam and Albert.

6.1 Pragmatism – William James

In order not to be accused of attacking a strawman, an interpretative issue needs first to be addressed.

For someone used to thinking of pragmatism as a doctrine that rejects the idea of truth as correspondence of propositions to reality, a reading of James can be quite startling.

For he not only seems to repeatedly endorse this doctrine, but emphasises that it is a necessary ingredient of the pragmatist conception of truth. For example, he states that:

> My account of truth is realistic, and follows the epistemological dualism of common sense. (James, 1978: 117).

> Truth is essentially a relation between two things, an idea, on the one hand, and a reality outside the idea, on the other. (1978: 91).

> Realities are not *true*, they *are*; and beliefs are true *of* them. (1978: 106).

> The notion of a reality independent of either of us, taken from ordinary social experience, lies at the base of the pragmatist definition of truth. (1978: 117).

Hence it would appear, if these statements are taken at face-value, as if James is endorsing the correspondence theory of truth. Or, as he calls it, the "intellectual theory". On this point there is some dispute. Thayer's introduction to the 1975 edition of *The Meaning of Truth* states that James did not reject the correspondence theory. Instead he was treating correspondence as a necessary condition of "pragmatic truth". The three conditions for "pragmatic truth" are, according to Thayer, that a belief must correspond to reality, cohere with other beliefs, and be beneficial[167] (1975: xxxvii).

This interpretation of James will not be followed here. Instead Ayer's interpretation in the 1978 introduction to a joint edition of *Pragmatism* and *The Meaning of Truth* will be used. There are two main reasons for doing so:

Firstly, if correspondence was meant to be a condition for pragmatic truth, then all disputes could have been avoided by simply using different terms to signify "correspondence" and "pragmatic truth". In fact, James dismisses the attempts of Hawtrey (1978: 316-318) and Pratt (1978: 90-98) to do this in order to clarify the issue.

---

[167] This interpretation followed by Hallberg (1997: 205-223) in a volume edited by Hardwick and Crosby.

Secondly, there is Ayer's argument for his interpretation, which appears sound. The whole argument will not be reconstrued, but the part that seems conclusive hinges on James' answer to Russell's challenge to his theory. Russell interprets James as dismissing correspondence, and claim that this implies that "the belief that A exists may be 'true' even when A does not exist" (1978: zzv). James states that this is "the usual slander, repeated to satiety by our critics" (1978: 147). However, Ayer doubts whether this is really slander, for James' argument against this seems to presuppose an unusual use of the word "true". Referring to the disputed authorship of "Shakespeare's plays", he states that:

> If the critic be both a pragmatist and a baconian, he will in his capacity of pragmatist see plainly that the workings of my opinion, I being what I am, make it perfectly true for me, while in his capacity of baconian he still believes that Shakespeare never wrote the plays in question (1978: 147).

As Ayer points out, the correspondence theory does not hold that truth is "for me", but precisely that it is independent of me:

> On this view, a belief cannot be true for one person and false for another, unless this is just a way of saying that one person may hold it and another not. If a belief is true there is a fact which makes it so; if it is false there is no such fact; and these facts obtain or fail to obtain, irrespectively of anyone's belief (1978: zzvi).

It would appear as if James' seeming endorsement of correspondence is a misleading one. The above criticism of Russell's objection invokes the idea that a person must *believe* a proposition to correspond to reality in order for it to be true, not that it *must so correspond*. And, as Ayer claims (1978: xxvi-xxviii), this distinction between corresponding to reality and being believed to correspond to reality might very well be exactly the distinction that James was arguing against. Whatever the merit of such a view may be, this is definitely not the correspondence theory of truth as understood by Russell, Ayer, and, I would contend, most philosophers.

Hence Ayer's interpretation of James will be followed. If this turns out to be false, at least the blow can be softened by knowing that anyone accused of misreading a pragmatist is in good company[168]. Turning to the substantive issue, what is James' claim?

James appears[169] to follow Peirce's pragmatic conception of meaning when he asks what truth's "cash-value" is "in experiential terms"(1978: 97). He answers that true ideas are those we can assimilate, validate, corroborate and verify. What is important to this analysis is that he connects these above concepts to truth as being "essentially bound up with the way in which one moment in our experience may lead us towards other moments which it will be *worth while* to have been lead to" (1978: 98, my italics.). Here truth is explicitly tied to the idea of being valuable. Indeed James makes the famous pronouncement that there is no difference between saying "it is true because it is useful" and "it is useful because it is true" (1978: 98).

We need not analyse his conception any further, the above statement is enough for us to proceed. What can the above possibly mean? Rejecting the idea that he was using "correspondence to reality" as a necessary condition for the above "truth", and remembering his endorsement of the idea that truth is "truth-for-someone", it can only be concluded that truth is "sacrificed" for value. The idea of "truth" or a "fact" as something conceptually distinct from "value" is simply inconsistent with statements like the above. Indeed, James concludes that the true "is only the expedient in the way of our thinking, just as 'the right' is only the expedient in the way of our behaving" (1978: 106).

Based on the interpretive matters discussed earlier, this means James can be treated as ultimately claiming that "truth" *is* "value" (the "expedient", the "worth while", "what works"). In other words, "truth" and "fact" can be fully reduced to value.

---

[168] *The Meaning of Truth* contains several essays that are rejections of criticism. Without fail, James admonishes his critics for misunderstanding him. Similarly Dewey (1939: 517-607), in an essay replying to *his* critics, states that he has "obviously failed... to make clear my actual position" (520), and proceeds to take his critics to task. Philosophers commonly claim misunderstanding on the part of their critics, but the pragmatist's repeated claim of systematic misunderstanding represents an extreme case. After reading several of these replies one might be forgiven for wondering whether any non-pragmatist has *ever* been admitted to fully understand pragmatism, and yet reject it.

[169] I say "appears" because Peirce was not always equally sanguine about James' use of his work. In fact, he changed the name of his theory of meaning to *pragmaticism*, "a name ugly enough to be safe from kidnappers." (Honderich, 1995: 709).

In chapter 5 it was explained why "truth" is a value, maybe the most important we have. If this is correct, then James can be credited with a great insight. James saw that truth is a value, and that there is no reason to hold beliefs in the first place if they do not serve human needs. He also noticed that truth is very valuable, i.e. that we cannot get along without the beliefs we hold to be true. Indeed, James' argument for the value of truth is strikingly similar to the ones employed here:

> The importance in human life of having true beliefs about matters of fact is a thing too notorious. We live in a world of realities that can be infinitely useful or infinitely harmful. Ideas which tell us which of them to expect count as the true ideas in all this primary sphere of verification, and the pursuit of such ideas is a primary human duty.... True ideas would never have been singled out as such, would never have acquired a class-name, least of all a name suggesting value, unless they had been useful from the outset in this way (1978: 98).

At face value this thesis can agree almost completely with the above paragraph. James is pointing out that true ideas fulfill a vital function in our lives, that if this wasn't the case we would have no need for true ideas. Indeed, he states that the determination of truth is a "primary human duty". In similar vein it was argued that the subject *should* try and attain the truth. This thesis agrees with everything except that he takes the above to imply that a conception of truth as conceptually distinct from value is nonsense. In chapter 5 it was shown that realising the value of valuable belief necessitates the conception of true belief as something different from valuable belief, but easier to determine. On this account, emphasising the "value" of truth is not something that needs to occur at the expense of the "truth" of truth. James need not have excluded a conceptually distinct notion of truth in order to affirm the value of truth. The two ideas are consistent.

A more extreme statement of some of the above sentiments can be found in Rorty.

6.2 Pragmatism - Richard Rorty

Richard Rorty explicitly rejects the claim that there are two distinct types of knowledge. Considering the claim that facts are cognitive and values not, he writes, in a discussion of Davidson, that:

> This picture is one that suggests that certain sentences in our language "correspond to reality" whereas others are true only, so to speak, by courtesy (1986: 1).

Rorty dismisses this view. The whole idea that sentences "correspond to reality" is, according to him, misguided. Indeed, he argues that the attempt to "explicate 'rationality' and 'objectivity' in terms of conditions of accurate representation is a self-deceptive effort to externalise the normal discourse of the day, and that... philosophy's self-image has been dominated by this attempt" (1980: 11). His argument for this can be described as neo-pragmatic. In support of it he enlists Kuhn, Davidson, Derrida, Sellars, Heidegger, James[170], Dewey, Wittgenstein, Gadamer[171] and others. Rorty's unqualified dismissal of the "correspondence theory" has the fortunate consequence that the interpretive problem incurred when looking at James does not occur. Indeed, *he himself interprets James* as having tried to provide "a utilitarian ethics of belief" (1997: 3). Rorty proceeds to defend and radicalise this view. In contrast to James' view, quoted above, that the seeking of truth is a "primary human duty", he approvingly interprets(!) James as claiming that:

> The view that there is no source of obligation save the claims of individual sentient beings entails that we have no responsibility to anything other than such beings. Most of the relevant sentient individuals are our fellow human beings. So talk about our responsibility to Truth, or to Reason, must be replaced by talk about our responsibility to our fellow human beings (1997: 3).

---

[170] It was already noted that James says some surprising things. Rorty is selective about what he considers to be "pragmatic" about James. As such Rorty calls James' statement that we are "recorders, not makers of the truth" "highly unpragmatic" (1997: 9).

Rorty's argument against the "correspondence" theory will, again, not form part of this discussion[172]. The claim relevant to this thesis is the claim that we do not have any responsibility to "Truth", but instead have a responsibility to "other beings".

The position taken in this thesis can enthusiastically affirm that we do, indeed have a responsibility towards "Truth". But, as was argued in chapter 5, "truth" is a "value". Hence the responsibility to determine the "truth" is not some metaphysical abstraction. Rather this "responsibility" is to our great benefit. One can reply to Rorty that our responsibility to "Truth" *is*, in fact, a responsibility to *ourselves*. On this account the idea that our only responsibility lies with sentient beings is no longer in conflict with the idea that we should determine the truth. "Responsibility", in the final analysis, does only apply to sentient beings, beings whose experience have a certain value.

Rorty's conception of our responsibility to truth as something distinct from responsibility to ourselves seems to rely on the type of idea criticised at the beginning of this chapter. This was the idea of "objectivity" as "value-free". If we connote "value-free" to "objective" then objective inquiry can no longer be portrayed as an act that is in the interest of human beings. Rorty appeals to this idea for criticism when he writes that:

> On the traditional account, desire should play no role in the fixation of belief. On the pragmatist account, the only point of having beliefs in the first place is to gratify desires (1978: 7).

On the account in this thesis these two aspects are not exclusive. It can be agreed that we only have beliefs in order to gratify desires, and agreed that they determine *what* we wish to have beliefs about. But, as pointed out, this does not erase the possibility of "objectivity". It does not erase the possibility that desire will play no role in the fixation (assigning of the predicate "true") of our beliefs.

---

[171] Rorty's interpretations of these philosophers has been questioned. Given the diverse collections of thinkers he marshals in support of his views this is probably not surprising. For a critique of his interpretation of Gadamer, see Warnke (1987: 139-166)

Rorty's dismissal of our responsibility towards truth does not realise that this responsibility is a responsibility towards ourselves. This seems to be the result of a false dichotomy. He supposes that inquiry must either be "for ourselves" or "for truth". But, as has been argued at length, this is a false dichotomy.

8. Conclusion

This chapter concerned some abstract epistemological matters that are implied everywhere in this thesis, but did not receive explicit attention. These concern the relation between fact and value, and the related conceptualisation of notions of objectivity. It was argued that, if the good is a *quale*, the relation between fact and value is that of a set to a subset.

This view has interesting implications for traditional views regarding objectivity. These views do not, however, rest on the specific view of the relation between fact and value that was presented in this thesis, but can be deduced from any view that allows them to be conceptually distinct. Since the issue regarding the intelligibility and coherence of the fact/value distinction is also a basic assumption of this thesis, objections to this distinction were considered in some detail. These considerations led to the conclusions that the objections are mistaken, and that there is no reason to abandon the distinction.

The idea that fact and value are distinct, coupled with the idea that a goal can be pursued by pursuing a logically independent, and sometimes conflicting, sub-goal, implies that what is important about the traditional ideal of objectivity is untouched by realising objectivity (and "truth") to be goals of actions, and therefore values. If fact and value are independent, then the fact that objectivity is a value cannot be used to doubt the coherence of claiming knowledge to refer to representation-independent entities, nor can it be used to doubt our ability to acknowledge these representation-

[172] With one exception. Rorty states that realists have to claim that reality has an intrinsic nature (1997: 5). Searle dismisses a variant of this argument as "remarkably feeble" (1998: 23). For a discussion see Searle (1998: 20-26).

independent entities where we would prefer that they did not exist. In other words, objectivity as epistemic value does not imply objectivity to be incoherent.

It was also argued that pragmatists often use the idea of truth as valuable to argue against the idea that truth relates to representation-independent entities. Hence the pragmatist conclusion, at least inasmuch as it is reached by this argument, is unwarranted.

The next chapter is the final one. It will attempt to clarify some issues that have been implicit thus far, without being specifically addressed. It will also serve as a conclusion to this thesis.

Chapter 8: Concluding Comments

1. Introduction

This chapter will try and briefly state the position, elsewhere mentioned only by implication, that this thesis takes concerning two matters that are clearly implicated in this attempt to formulate the outlines of a cognitive ethics. The first issue concerns the problem regarding actions that greatly conflict with our moral intuitions. The second concerns the merit of the whole idea of having a cognitive theory of ethics. These discussions of the gains from such a view will also serve as a conclusion to this thesis.

2. The Problem of Actions that Greatly Conflict with Our Moral Intuitions

2.1 The Problem and Whether Ethical Hedonism Solves It

How can an egoist ethics deal with cases where people commit actions that greatly conflict with our moral intuitions? If someone was to, for instance, take great pleasure in killing and robbing, is their action thereby justified?

If someone commits acts that greatly conflict with our moral intuitions there can be two senses in which an Egoist can "condemn" these actions. The first is by pointing out that the person "should" not have committed these actions, i.e. they cannot serve his long-term self-interest. Hence it can be said that these actions were based on a mistake of the type discussed above, and therefore were mistakenly judged to be in a person's self-interest.

But what about the person who enjoys killing or robbing, won't suffer pangs of conscience, and might well not be caught - a psychopath? How can one counter the action of the Hobbesian "foole", or Hume's "sensible knave"? Here there exists an instance where self-interests conflict, either because the person is robbing or killing me, or because he is robbing someone in whose life I have a positive interest, or because I have hatred of robbing and killing, etc. Here this thesis has nothing, in principle, to offer over and above portraying such a situation as a clash of wills. The

other has an interest in robbing and killing, and I have an interest in stopping him. It seems that the issue will be decided based on our relative power[173].

The majority of society can foresee this type of possibility, and see that it is in their own individual interests to create systems of law that increase the penalty attached to such action, thereby making its occurrence more unlikely. It can also see that it is in the interests of the majority of society to remove such people from society, i.e. prison can be used as a type of assortative mechanism. But the person cannot be "condemned" in the sense that action is "condemned" within, for instance, a deontological ethical system. Anscombe's classic paper *Modern Moral Philosophy* (1981) convincingly argues that our moral categories are attached to worldviews that a lot of people no longer accept[174]. The case of "condemnation" seems to be such a case if ethical hedonism is accepted; certain moral categories perish when certain worldviews perish.

Hence ethical hedonism does result in certain cases where action that radically conflicts with our ethical intuitions cannot be "condemned". It does not, of course, imply that we must "condone" such action; both categories are jettisoned by a hedonist ethics. If it maximises value then we should act so as to prevent such action. But ethical egoism cannot go beyond treating such cases as clashes of will.

Hence an egoist ethics cannot help when dealing with the "sensible knave" that was the source of much trouble for Hume. It could possibly argue that there would be very few sensible knaves because of evolutionary pressures against such actions. It can definitely argue that those who are troubled by it should oppose it. But it cannot, in principle, condemn any action, apart from showing that it is unrelated to self-interest.

---

[173] The above clash has been referred to as a "paradox of egoism", and has been used to call egoism incoherent. For a discussion of why this is not the case, see Singer (1993: 319-320).

[174] "[T]he concepts of obligation, and duty - *moral* obligation and *moral* duty, that is to say – and of what is morally right or wrong, and of the *moral* sense of "ought", ought to be jettisoned if this is psychologically possible; because they are survivals, or derivatives from survivals, from an earlier conception of ethics which no longer generally survives, and are only harmful without it" (Anscombe, 1981: 27).

2.2 Relevance of the Above Problem

It is important to be clear about what is being conceded above. It is being conceded that an egoist hedonist ethics has one very counter-intuitive implication. It is not being conceded that this should count against the validity of cognitive, egoist, ethical hedonism.

It was already remarked, in chapter 6 and in connection with altruism, that the validity of a cognitive ethics cannot be undermined by referring to any action that such an ethics does or does not recommend. The validity of the egoist hedonism defended in this thesis depends entirely on whether the claim that the good is a *quale* is accepted. Arguments as to the validity of egoist hedonism need to attack the validity of this conceptualisation of the good.

If one has an epistemological justification for treating a strong intuition as part of what needs to be counted when it is determined whether one is justified in believing something, then the above problem is one factor that counts against the rationality of accepting egoist hedonism. But this does not relate to the *validity* of egoist hedonism, rather it relates to whether a belief in egoist hedonism is a *justified belief*. In other words the situation is similar to one where an authority on some matter tells us his view of the matter. To treat the person's authority as conclusive proof of the truth of his views is to commit the logical fallacy of appeal to authority. But the fact that the person is an authority does make it more rational to accept what he says. In a similar manner the problem outlined above gives us grounds for being suspicious of egoist hedonism, and reason to investigate its claims with great care. But it does not offer conclusive proof of the fallacious nature of egoist hedonism.

3. The Gains Offered by a Cognitive Theory of Ethics

In chapter 1 it was argued that, given the categories of folk psychology and the existence of *qualia*, a coherent meta-ethical position that allows for cognitive statements regarding value can be formulated. It was argued that the folk

psychological view that we are actors with certain irreducible desires can, on its own terms, be shown to be incomplete. The idea of a desire or preference as a "fundamental want" cannot be the final truth concerning human action. This is because it cannot state why a certain object is wanted, rather than another. This explanatory gap can only be closed if some element of our experience enabled the regress of justification to end in a non-arbitrary manner. This regress can only be ended if there is something that, ultimately, simply is *good*.

It was argued that the whole idea of this regress ending in a non-arbitrary manner only makes sense if we consider the phenomenon that we often call "pleasure". While the word "pleasure" has a great many uses over and above referring to this self-justifying element of experience, it was argued that there are uses of this term where it refers to this self-justifying element of our experience. This self-justifying element was called "value", and the main claim of this thesis is that it forms part of the inventory of entities we encounter in the world.

This entity was then characterised as a phenomenal quality or *quale*, and hence part of the first person, subjective ontology of the world. If the characterisation of value as *quale* is accepted, it is immediately evident that this implies a form of egoism. The sense in which "pleasure" is self-justifying is not shared, and hence the pursuit of "pleasure" is irrevocably tied to a specific individual's experience.

To say that this entity is good is equivalent to saying that it should be pursued. Hence the characterisation of the object of human motivation implies ethical hedonism, and not psychological hedonism[175]. If this is accepted, then it follows that the study of ethics is the study of how individuals should maximise pleasure. The specifics as to how this is to be done is a very complicated question.

---

[175] Does the doctrine of egoist hedonism have any specific implication regarding the "freedom of the will"? This topic is one of those that seem to be a problem for most explanations of human action. For can action capable of explanation also be free? Philosophers have made ingenious attempts to save apparently deterministic views of man from this implication. This is not a topic that can fruitfully be discussed in this thesis. I do, however, wish to note one consequence of adopting ethical, as opposed to psychological, hedonism. If ethical hedonism does result in determinism, and it is not sure that it does, then this determinism might well be peculiar in one way. If ethical hedonism is true, then a being with free-will and no cognitive constraints would choose the most pleasurable option any time, simply because of the special nature of pleasure. This means that pleasure is chosen because it *should* be chosen, not because one is in some sense *forced* to. In other words it is not that one *cannot* choose the less pleasurable option, rather one simply *does not* do so.

One specific part of it was addressed when it was argued that, subject to certain qualifications, people should treat "truth" as a norm of inquiry. It was also claimed that it is not at all clear that an egoist hedonist would *not* commit any number of actions that we normally consider to be altruistic. This was done in order to cast doubt on the intuitive plausibility of thinking that egoist hedonists would necessarily be "vicious" and "self-serving", as these terms are commonly understood when used to condemn certain actions.

The gains from the above position vest in the fact that it would make ethical statements continuous with ordinary descriptive statements. In other words we can know that there is a fact of the matter concerning how we should act. It might be thought that this truth is an unappealing one, and that the guide offered by other theories of ethics is preferable. But here egoist ethics has the one advantage in that justifying the use of this guide is, in principle, a simple matter. We should act as it prescribed because it is true, i.e. reality advises us to act in this way. Non-cognitive theories ultimately face the difficulty of justifying their fundamental claims. Here one option is to simply say that there are two distinct criteria for justifying statements, i.e. one set for descriptive statements and one set for ethical statements. Alternatively it can disregard purely descriptive statements, as pragmatism, on some interpretations, seems to, or disregard ethical statements, as positivism seems to. The above options seem entirely unappealing.

In providing a standard against which to measure action, ethical hedonism also enables us to make sense of the idea that people can act against their own best interests. We live in a time where a lot of people are deeply distrustful of any grand justification of action. Yet the vacuum left by the loss of belief in ethical systems is not a vacuum that can go unfilled. In some way or other life calls upon us to judge and act. Here a sort of "default ethics" has emerged, whereby people should be treated as beings that need to find their own way as best as possible, and should not be criticised in terms of ethical views that are imposed upon them from outside sources. This type of "default ethics" of tolerance, while not without considerable virtues, has certain limitations. The first is that it cannot be a guide for action, except inasmuch as it outlaws intolerance. But the majority of our actions have nothing to do with

tolerance or intolerance. Rather we have to choose between any number of options where there is nothing clearly tolerant or intolerant about any of them.

The second problem is that it cannot always be used to condemn things in our society that we might wish to. If any action that does not in any obvious way interfere with the rights of others, and is in some sense the outcome of choice, is thereby also legitimate, then the quietist consequences cannot be avoided. Criticism is often seen as paternalistic infringements of autonomy, and therefore dismissed. This type of situation can only be averted if there is a standard whereby action can be judged, so that it makes sense to say to someone that he has acted against his own best interests.

This type of criticism is possible if the fundamental claim of this thesis is accepted. It is possible for two related reasons; the first is the obvious one that it gives a hedonist standard against which actions can be judged. The second reason is more subtle, and concerns the view of man as boundedly rational rule-follower. If rational choice theory is an accurate description of human decision-making, and all action is a function of desires and beliefs, then one can only act against one's own interests[176] because of inadequate information. The actual mechanism whereby decisions are made can never be part of the explanation for why someone has acted against his own interests. This makes it hard to see how systematically acting against one's interests is possible.

The situation changes if one views people as boundedly rational. Bounded rationality is context specific, and the rules according to which it operates can be extremely unsuccessful in unfamiliar contexts. Hence acting against one's own interests can be systematic if the right context, one that exploits the limitations in a boundedly rational rule, is generated. This means that a *prima facie* case can be made against actions that are only committed due to weaknesses inherent to boundedly rational rules[177]. In other words bounded rationality has less quietist consequences than rational choice.

---

[176] This assumes of course, that one's desires should be treated as having some normative import, which is not itself a part of any version of rational choice theory.

[177] For an example of such a boundedly rational rule that can be to the detriment of the actor, see Ainslie's *Breakdown of Will* (2001). It is argued that weakness of will can be explained by seeing the agent's discount curves as hyperbolic, rather than exponential. This leads to inconsistent preferences over time. Such inconsistent preferences can lead to being exploited *via* "money-pumps", and yet might have been adaptive in our evolutionary past (2001: 45-47).

The above point is not made in order to justify wholescale interference with human action. It might well be that the attempted cure is a more dangerous poison than the problem itself. But the situation can perhaps be improved by something a long way from the nightmare visions of totalitarianism that we are so eager to associate with any criticism of individual choice. Something as simple as educating someone about his bounded rationality, as it concerns a specific action, could possibly have the desired effect of changing action. Here it is a case of letting someone see how the unwanted effect is generated. The effect can hereby lose its power, as it is unlikely that the rules that will relate this new information to action are subject to the same effect. It might lose its power in the same way that a magician's stage effect loses its power when we realise how it is done.

Specific study as to specific action-generating rules, and how they cause behaviour will be immensely useful here. Something like, for instance, the lament as to the culture industry's dumbing down of culture can become more than a seemingly elitist imposing of different views if it can be shown that the culture industry exploits certain specific decision-making rules that might have been adaptive for a hunter-gatherer, but are counterproductive in modern society. Here ethical hedonism adds the final touch by explaining why, and in what sense, our desires have moral import.

The view presented in this thesis provides one way of thinking about the type of critique discussed above. Two *prima facie* likely phenomena that can be exploited to cause one to act against one's own interests are "weakness of will" (akrasia) and self-deception. They are phenomena that cannot possibly occur within the rational choice view of human action. But bounded rationality, coupled with some degree of realism concerning preferences, provide an avenue whereby they can be coherently be explained. If action can be the result of both boundedly rational rules, chosen mostly by evolution in order to enhance fitness, and preference maximisation, which is only related to evolution as secondary consequence, then the actor is effectively split into two actors. One is trying to maximise a set of preferences, while the other actor is, *metaphorically speaking*, trying to manipulate the first into maximising fitness. A "multiple selves" view of action has often been thought to be exactly what is required

in order to explain phenomena like akrasia and self-deception[178]. Something like self-deception, for instance, need in principle present no more conceptual difficulty than the case from ordinary life where one person lies to another. This avenue of acquiring these conceptual gains cannot be had unless there is a difference, in principle, between action that is in some sense the result of a preference, and "action" which is not. This is not possible if action is defined only in terms of behaviour, and the situation is left at that.

Hence I wish to argue that egoist hedonism can avoid the quietist consequences of the "default view", both because of cognitivity and because of the view of man as boundedly rational actor.

I would like to end this thesis by citing a goal that, as argued in chapter 5, is often taken to be such a self-evident goal that it is not realised that it is a goal. The main gain of trying to formulate a cognitive ethics lies in the fact that success would imply that one can have value-judgements that are *true*. And, as argued in chapter 5 and chapter 7, this is one of the most important and useful values we have.

[178] See, for example, the collection *The Multiple Self* (1987), edited by Jon Elster, and Ainslie (2001).

Bibliography

Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

Albert, H. 1985. *Treatise on Critical Reason*. New Jersey: Princeton University Press.

Anscombe, GEM. 1981. *The Collected Philosophical Papers of GEM Anscombe. Volume 3: Ethics, Politics and Religion*. Oxford: Blackwell.

Austin, JL. 1964. *Sense and Sensibilia*. Oxford University Press: Oxford.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. 1997. *The Complexity of Cooperation*. Princeton University Press: Princeton, New Jersey.

Ayer, A. J. 1982. *Language, Truth and Logic*. Harmondsworth: Penguin Books.

Ayer, A.J. 1985. *Wittgenstein*. London: Weidenfeld and Nicolson.

Badcock, C. 2000. *Evolutionary Psychology: A Critical Introduction*. Polity Press: Cambridge.

Barkow, J, Cosmides, L & Tooby, J. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Becker, G.S. 1976. *The Economic Approach to Human Behavior*. University of Chicago Press: Chicago.

Beer, FA. 1986. "Games and Metaphors" in *Journal of Conflict Resolution*, 30 (1): 171-191.

Bentham, J. 1982. *An Introduction to the Principles of Morals and Legislation*. London: Methuen.

Binmore, K. 1994. *Game Theory and the Social Contract. Vol1: Playing Fair*. MIT Press: Cambridge.

Binmore, K. 1998. *Game Theory and the Social Contract. Vol 2: Just Playing*. MIT Press: Cambridge.

Bohm-Bawerk, E. 1959. *Capital and Interest. Volume 3*. South Holland: Libertarian.

Brodie, AB. (ed). 1970. *Readings in the Philosophy of Science*. New Jersey: Prentice-Hall Inc.

Broome, J. 1994. "The Structure of Good: Decision Theory and Ethics" in Bacharach, M & Hurley, S. (eds). *Foundations of Decision Theory*. Oxford: Blackwell.

Burwood, S, Gilbert, P & Lennon, K. 1998. *Philosophy of Mind*. London: UCL Press.

Campbell, T. 1981. *Seven Theories of Human Society*. Oxford: Clarendon Press.

Cilliers, P. 1998. *Complexity and Postmodernity*. London: Routledge.

Churchland, PM. 1994. "Folk Psychology (2)" in Guttenplan, S. (ed.) 1994. *A Companion to the Philosophy of Mind*. Oxford: Blackwell.

Cornell, D. 1992. *The Philosophy of the Limit*. New York: Routledge.

Dennett, D. 1987. *The Intentional Stance*. Cambridge: MIT Press.

Eatwell, J, Milgate, M & Newman, P.(eds.) 1987. *The New Palgrave: Marxian Economics*. London: Macmillan Press.

Elster, J. 1985 (ed). *The Multiple Self*. Cambridge: Cambridge University Press.

Elster, J.(ed). 1986. *Rational Choice*. New York: New York University Press.

Elster, J. 1990a. *The Cement of Society*. Cambridge: Cambridge University Press.

Elster, J. 1990b. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Frank, R.H. 1990. "A Theory of Moral Sentiments" in Mansbridge, J.J. *Beyond Self-Interest*. University of Chicago Press: Chicago.

Frankena, WK and Granrose, JT. (eds). 1974. *Introductory Readings in Ethics*. New Jersey: Prentice-Hall Inc.

Gelven, M. 1970. *A Commentary on Heidegger's "Being and Time"*. New York: Harper & Row.

Gigerenzer, G, Todd, M & The ABC Research Group. 1999. *Simple Heuristics that Make Us Smart*. Oxford University Press: Oxford.

Gigerenzer, G & Selten, R. (eds). 2002. *Bounded Rationality: The Adaptive Toolbox*. MIT Press: Cambridge.

Goddard, L & Judge, B. 1982. The Metaphysics of Wittgenstein's 'Tractatus', in *The Australasian Journal of Philosophy*. Monograph 1, June 1982.

Guttenplan, S. (ed.) 1994. *A Companion to the Philosophy of Mind*. Oxford: Blackwell.

Hallberg, FW, 1997. "Coherence, Correspondence, Satisfaction, Power: The Four Elements of James' Pragmatic Theory of Truth" in Hardwick, CD and Crosby, DA (ed's). *Pragmatism, Neo-pragmatism and Religion: Conversations with Richard Rorty*. New York: Peter Lang.

Hamilton, WD. 1963. "The Evolution of Altruistic Behaviour" in *American Naturalist*. Vol 97: 354-356.

Harris, CE, Pritchard, MS and Rabins, MJ. 2000. *Engineering Ethics: Concepts and Cases*. Belmont: Wadsworth Publishing Company.

Heidegger, M. 1987. *Being and Time*. Southampton: Blackwell.

Held, D. 1983. *Introduction to Critical Theory: Horkheimer to Habermas*. London: Hutchinson and Co.

Hicks, JR. 1962. *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*. Oxford: Clarendon Press.

Hobbes, T. 1973. *Leviathan*. London: JM Dent & Sons.

Hoffrage, U & Hertwig, R. 1999. "Hindsight Bias: A Price Worth Paying for Fast and Frugal Memory" in Gigerenzer, G, Todd, M & The ABC Research Group. 1999. *Simple Heuristics that Make Us Smart*. Oxford University Press: Oxford.

Honderich, T. (ed). *The Oxford Companion to Philosophy*. Oxford: Oxford University Press.

Hume, D. 1969. *A Treatise of Human Nature*. Harmondsworth: Penguin Books.

Hutchison, T. 2000. *On the Methodology of Economics and the Formalist Revolution*. Edward Elgar: Cheltenham.

James, W. 1975. *Pragmatism* and *The Meaning of Truth*. Cambridge: Harvard University Press.

James, W. 1978. *The Meaning of Truth*. Cambridge: Harvard University Press.

Jevons, W.S. 1911: *The Theory of Political Economy.* London: Macmillan and Co.

Joyce, R. 2001. *The Myth of Morality*. Cambridge: Cambridge University press.

Kant, I. 1964: *Groundwork of the Metaphysics of Morals*. New York: Harper and Row.

Kauder, E. 1965: *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Klein, G. 2002. "The Fiction of Optimisation" in Gigerenzer, G & Selten, R. (eds). 2002. *Bounded Rationality: The Adaptive Toolbox*. MIT Press: Cambridge.

Koertge, N. (ed). 1998. *A House Built on Sand: Exposing Postmodernist Myths About Science*. Oxford: Oxford University Press.

Layard, R. 2003. *Happiness: Has Social Science a Clue?* Lionel Robbins Memorial Lectures 2002/3.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin Books.

Marx, K. 1976. *Capital: Volume 1*. Harmondsworth: Penguin Books.

Mclelland, D. (ed) 1977. *Selected writings of Marx.* Oxford: Oxford University Press.

Moore, G.E. 1968. *Principia Ethica*. London: Cambridge University Press.

Monroe, K. R. 2001. Paradigm Shift: From Rational Choice to Perspective. in *International Political Science Review*, 22 (2), 151-172.

Nagel, T. 1981 [1974]. "What is it Like to be a Bat?" in Hofstadter, D.R. and Dennet, D.C. (eds.). 1981. *The Mind's I: Fantasies and Reflections on Self and Soul*. London: Penguin Books.

O' Brien, D. P. (1975): *The Classical Economists.* Oxford: Clarendon Press.

Pigden, CR. 1993. "Naturalism" in Singer, P (ed). 1993. *A Companion to Ethics*. Oxford: Blackwell.

Popper, KR. 1976. *Unended Quest: An Intellectual Biography*. Glasgow: Fontana Collins.

Putnam, H.1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, WV. 1965. "Two Dogmas of Empiricism" in Ammerman, RR. 1965. *Classics of Analytic philosophy*. New York: Mcgraw-Hill, Inc.

Rand, A. 1989. *The Voice of Reason*. New York: Meridian.

Rescher, N. 1969. *Introduction to Value Theory*. New Jersey: Prentice Hall.

Ridley, M. 1996. *The Origins of Virtue*. New York: Penguin Books.

Rorty, R. 1980. *Philosophy and the Mirror of Nature*. Cambridge: Blackwell.

Rorty, R. 1986. Unpublished Paper.

Rorty, R. 1997. "Religious Faith, Intellectual Responsibility and Romance" in Hardwick, CD and Crosby, DA (ed's). 1997. *Pragmatism, Neo-pragmatism and Religion: Conversations with Richard Rorty*. New York: Peter Lang.

Ross, D. 1999. *What People Want: The Concept of Utility from Bentham to Game Theory*. Cape Town: UCT Press

Ross, D. 2002. "Why People are Atypical Agents" in *Philosophical Papers*. 31: 87-116.

Ruse, M. 1986. *Taking Darwin Seriously*. New York: Blackwell.

Ruse, M and Hull, DL. (eds.) 1998. *The Philosophy of Biology*. Oxford: Oxford University Press.

Samuelson, P. and Nordhaus, W. D. (1989): *Economics (thirteenth edition).* New York: Mcgraw-Hill Book Co.

Sartre, J-P. 1981. *Being and Nothingness*. London: Methuen and Co.

Schillp, PA. (ed). 1939. *The Library of Living Philosophers. Volume 1: The Philosophy of John Dewey*. Chicago: Northwestern.

Schlick, M. 1962. *Problems of Ethics*. New York: Dover Publications.

Schumpeter, J.A.1994. *History of Economic Analysis*. London: Routledge

Searle, J.R. 1992. *The Rediscovery of the Mind*. Massachusetts: MIT Press.

Searle, J.R. 1998. *Mind, Language and Society: Philosophy in the Real World*. New York: Basic Books.

Sedgwick, H. 1901. *The Methods of Ethics*. London: Macmillan and Co.

Shepard, RN. 1992. "The Perceptual Organisation of Colours: An Adaptation to Regularities of the Perceptual World?" in Barkow, J, Cosmides, L & Tooby, J. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Simon, H. 1997. *Models of Bounded Rationality. Volume 3: Empirically Grounded Economic Reason*. Cambridge: MIT Press.

Simon, H. 1997. *An Empirically Based Microeconomics*. Cambridge: Cambridge University Press.

Singer, P. 1993. *Practical Ethics* (2nd ed). New York: Cambridge University Press.

Smith, A. 1975. *The Wealth of Nations.* London: JM Dent and Sons.

Smith, H. 1994. "Deciding How to Decide: Is There a Regress Problem?" in Bacharach, M & Hurley, S. (eds). *Foundations of Decision Theory*. Oxford: Blackwell.

Sober, E. 1998. "What is Evolutionary Altruism?" in Ruse, M and Hull, DL. (eds.) 1998. *The Philosophy of Biology*. Oxford: Oxford University Press.

Sober, E. and Wilson, DS. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behaviour. Massachusetts: Harvard University Press.*

Sprigge, T. L. 1990. *The Rational Foundations of Ethics*. London: Routledge.

Stigler, J.G & Becker, G.S. 1967. "De Gustibus Non Est Disputandum" in *The American Economic Review*. 67: 76-90.

Tooby, J & Cosmides, L. 1992. "Cognitive Adaptations for Social Exchange" in Barkow, J, Cosmides, L & Tooby, J. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Trivers, R. 1971. "The Evolution of Reciprocal Altruism" in *Quarterly Review of Biology*, vol 46: 35-57.

Ullman-Margalit, E. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.

Waismann, F. 1979. *Wittgenstein and the Vienna Circle: Conversations Recorded by Friedrich Waismann*. (ed: B. McGuiness). Oxford: Blackwell.

Warnke, G. 1987. *Gadamer: Hermeneutics, Tradition and Reason*. Oxford: Blackwell.

Wilson, DS. 1998. "On the Relationship Between Evolutionary and Psychological Definitions of Altruism and Selfishness" in Ruse, M and Hull, DL. (eds.) 1998. *The Philosophy of Biology*. Oxford: Oxford University Press.

Wittgenstein, L. 1960. *Tractatus Logico-Philosophicus*. London: Routledge.

Wittgenstein, L. 1980. *Wittgenstein's Lectures, Cambridge, 1930-1932*. Oxford: Blackwell.

Wittgenstein, L. 1984. *Culture and Value*. Chicago: University of Chicago Press.

Wittgenstein, L. 1994. *Philosophical Investigations*. Oxford: Blackwell.