ACTA PHILOSOPHICA GOTHOBURGENSIA
6

The Human Good

Bengt Brülde

ACTA UNIVERSITATIS GOTHOBURGENSIS

ACTA PHILOSOPHICA GOTHOBURGENSIA
6

The Human Good

Bengt Brülde

AKADEMISK AVHANDLING
för avläggande av filosofie doktorsexamen i praktisk filosofi, som med tillstånd av humanistiska fakultetsnämnden vid Göteborgs universitet framläggs till offentlig granskning lördagen den 14 februari klockan 12 i Stora Hörsalen, Humanisthuset, Renströmsgatan 6, Göteborg.

Abstract

Title: The Human Good (Acta Philosophica Gothoburgensia no 6)
ISSN 0283-2380, ISBN 91-7346-324-8
Language: English text, with a summary in English
Author: Bengt Brülde
Doctoral dissertation at the Department of Philosophy, Göteborg University

The purpose of this thesis is to find the most plausible answers to the following three central questions of prudential value (or well-being): (I) What does a person's well-being consist in: what has final value for a person? (II) How do we determine just how valuable a certain situation (or fact) is for a certain person? And (III) how do we determine how well off a person is on the whole (at a certain time)?

In order to achieve this purpose, I conduct a critical examination of three common types of answers which have been given to these questions, viz. hedonistic theories, desire theories, and "objective list theories". For each of these theories of prudential value, I first try to offer a formulation of the theory which is as precise as possible (and which makes the theory as plausible as possible). I then try to find out whether the theory in question is a plausible theory, by looking at a number of arguments that can be given for and against the theory.

My conclusions can be formulated as follows: On the negative side, all three theories examined suffer from certain defects; especially the pure forms of these theories. On the positive side, I think it is possible to construct a theory that can avoid the objections which hit the other theories. This theory is a type of mixed theory: It can (roughly) be regarded as a mixture between modified hedonism and a certain type of modified actual desire theory, a mixture which also contains certain "objectivist" elements.

Key words
Well-being, welfare, good life, quality of life, human good, prudential value, value-for, theory of value, subjectivism, objectivism, value and time, hedonism, desire theories, preferentialism, objective list theory, pleasure, preference, rational desire

The Human Good

ACTA PHILOSOPHICA GOTHOBURGENSIA
6

The Human Good

Bengt Brülde

ACTA UNIVERSITATIS GOTHOBURGENSIS

Distribution:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222
SE-405 30 Göteborg
Sweden

ISBN 91-7346-324-8
ISSN 0283-2380

Printed in Sweden by
Kompendiet, Göteborg 1998

Acknowledgements

My first acknowledgement is to Björn Haglund, my supervisor, for his support and for our stimulating conversations.

Next, I want to thank all those who have read and commented on various parts of the manuscript. In particular my gratitude goes to Krister Bykvist, Mats Furberg, Anders Tolland, Torbjörn Tännsjö, and Jan Österberg.

I am also grateful to the practical philosophers at Uppsala University, for their helpful comments and encouragement. In particular I want to thank Krister Bykvist (again), Erik Carlson, and Jan Österberg (again), for our inspiring discussions on the desire theory.

Finally, I would also like to my gratitude to the Fulbright Commission, who made it possible for me to spend the year 1992/93 in New York.

A general note to the reader

As John Rawls once said (in *A Theory of Justice*), "[t]his is a long book, not only in pages". Earlier drafts (which also included a number of sections on the issue of measurability) were even longer, though. The reason why the book is so long is that it is problem-oriented in a very strict sense: I have been doing what I consider necessary in order to come up with satisfactory answers to the questions asked.

Now, I realize that there are a number of possible readers who will regard the present book as too long (and "not only in pages"). To make things easier for those people, I want to give "a few remarks by way of guidance". The central questions of this book are formulated on pp 22-24. The most fundamental passages besides this one are section 1.2, where the traditional theories of prudential value (viz. hedonism, the desire theories, and the objective list theory) are formulated, and chapter 8, where my own mixed theory of prudential value is presented. Those who want more precise formulations of the traditional theories can turn to chapters 2 (hedonism), chapter 4 (the unrestricted desire theory), section 5.2 (the modified desire theories), and chapter 6 (the objective list theory). Those who are interested in the arguments that can be given for and against these theories can turn to chapter 3 (hedonism), section 5.1 (the unrestricted desire theory), sections 5.2 and 5.3 (the modified desire theories), and chapter 7 (the objective list theory).

CONTENTS

CHAPTER ONE: The questions, the traditional answers, and some notes on "method" 1

1.1. The questions of prudential value 3
Goodness-for-people: What it essentially consists in 3
Value-for vs. value-period 9
A note on Betterness-for 11
The notion of well-being 13
Well-being and time: the temporal aspect of (III) 16
Lives are wholes 17
Evaluative atomism vs. evaluative holism 19
The problem of (synchronic) aggregation 21
The questions of prudential value revisited 22
A note on Normative relevance 25

1.2. A brief overview of the main answers to (I)-(III) 26
Substantive vs. formal theories of prudential value 26
The Hedonistic Theory 29
The satisfaction interpretation of the Desire-Fulfilment Theory 35
The desire theory vs. hedonism 41
"The Objective List Theory", or "non-internalist pluralism" 45
Monism vs. pluralism 45
Internalism, relationalism, and externalism 47
(i) Internalism 49
(ii) Relationalism 50
(iii) Person-Externalism 51
(iv) Life-externalism 51
The Objective List Theory revisited 52

1.3. On "method": Some epistemological assumptions 56
My epistemological assumptions: Weak cognitivism 56
In what general way (or ways) can value-for-claims be justified (or refuted)? 60

CHAPTER TWO: A formulation of the hedonistic theory (and its different versions) 68

2.1. To what exactly is pleasantness and unpleasantness primarily attributed? 73
Some preliminaries: How we attribute pleasantness and unpleasantness in ordinary speech 73
Pleasant and unpleasant sensations 76
Pleasant and unpleasant emotions 76
Pleasant and unpleasant moods 77
Pleasant experiences vs. pleasant "experience" 79

2.2. What do the terms "pleasantness" and "unpleasantness" refer to? Different conceptions of pleasantness and unpleasantness 81
Two kinds of conceptions of pleasantness (unpleasantness) 83
The Quality Theories 84
The Relational Theory 86
Three different versions of pure hedonism 91
The Quality Hedonisms 91
Preference-Hedonism 93
A critical discussion of the different conceptions 95
The Quality Theory vs. the Relational Theory: A general discussion 96
A heavy objection to the Monistic Quality Theory 100
Two objections against the Pluralistic Quality Theory 102
Why we should prefer the Relational Theory to the Pluralistic Quality Theory 103

CHAPTER THREE: Is hedonism a plausible theory of prudential value? A critical discussion of the hedonistic theory 105

3.1. Arguments for hedonism 107
Arguing for (justifying) hedonism: some general remarks 107

The three most common types of pro-arguments 108
On subject-oriented vs. object-oriented justification of value-for-statements, and why subject-oriented justification is necessary 110
Subject-oriented arguments for hedonism 113
Mill's argument 113
Wetterström's argument 115
Appealing to hypothetical desires or evaluations 119

3.2. Arguments against hedonism 120
Arguing against (criticizing, refuting) hedonism: some general remarks 120
The arguments against hedonism: A brief overview ...124
Arguments directed against (H1) and/or (H2) 125
Arguments which purport to show that a person's welfare can be directly affected by things he doesn't know anything about 126
A special case: Post-mortem events 127
Arguments that purport to show that the desire theory is superior to hedonism 128
A general argument against the experience requirement 128
Is death an evil for the person who dies? 129
Arguments that purport to show that pleasure is not all that matters 132
Can pleasant sensation ever be sufficient to make a life good? Nagel's contented infant 132
Griffin's argument 135
Rachels' argument 136
Nozick's "Experience Machine" 137
A short note on Mill's Pig 140
HAPPY 141
Arguments against pure hedonism that are, at the same time, arguments for some modified version of the hedonistic theory: The criticism of pleasures 142
Conclusion 152

CHAPTER FOUR: A formulation of the Unrestricted Desire-Fulfilment Theory 154

4.1. What is a desire? 164
The desiring subject 164
The object of desire: *desire for* and *desire that* 165
The thing-view 166
The situation-view: Desires as propositional attitudes 166
Why the situation-view is superior in ethical contexts 168
The desiring (by the subject, of the object): A rudimentary conception of desire 170

4.2. What is it for a desire to be stronger than another desire? 172

4.3. Desire and time: Some temporal issues 176

4.4. The intrinsicality condition 182

CHAPTER FIVE: Is the desire theory (in any of its versions) a plausible theory of prudential value? A critical discussion of the desire theories 187

5.1. A critical discussion of the unrestricted desire theory ... 189

5.1.1. Arguments for the unrestricted desire theory 190
"Arguments" for (D1) and against other theories 193
Arguments for (UD2)-(UD3) and against modified desire theories 195

5.1.2. Arguments against the unrestricted desire theory 197
Arguing against the unrestricted desire theory: A few general remarks 197
The arguments against the unrestricted desire theory: A brief overview 200
Object-oriented objections to the unrestricted actual desire theory 205

Rationality-oriented objections to the unrestricted actual desire theory 207
Other types of objections to the unrestricted actual desire theory (objections that are neither object-oriented nor rationality-oriented) 211
Evaluative objections to (UD4) 214

5.2. How the unrestricted desire theory can be modified in order to handle the objections above: Different kinds of modified versions of the actual desire theory 219

5.2.1. Some object-oriented modifications of the unrestricted desire theory 223
The Success Theory 224
About the temporal boundaries of a life 226
About the "atemporal" boundaries of a life 228
My life and the lives of (significant) others 228
My life and the "lives" of the wholes with which I identify, or to which I belong 230
The experience-oriented Success Theory 231
The Global Success Theory 232

5.2.2. Rationality-oriented modifications of the Success Theory 236
Rationality-oriented modifications based on Hume's theory 239
Griffin's informed desire theory 242
Rationality-oriented modifications based on the deliberative theory 245
Rationality-oriented modifications based on the genetic theory 253
Rationality-oriented modifications based on the intrinsic theory 257
Conclusion 259

5.2.3. Some other possible modifications of the (actual) desire theory 261
The appeal to evaluations and higher-order desires 261
Knowledge-oriented modifications 266
My conclusion: The most plausible version of the

actual desire theory 271

5.3. Is any version of the actual desire theory plausible? 272
Is there any truth in the idea that hypothetical desires should count? 275
Is there any truth in "the objective list theory"? 278
Scanlon's argument 279

CHAPTER SIX: "The objective list theory" I: A list of possible prudential values 286

A list of possible prudential values 288
(1) Activities and other "agent-goods" 289
(2) Social and relational goods 290
The special case of filia: its nature and its forms. 291
(3) Experiences and other mental states 294
(4) To be (qua experiencing and thinking subject) in contact with reality 294
(5) The prudential value of being a certain kind of person and/or living one's life in a certain way 295
(5:1) Autonomy 298
(5:2) Functioning in accordance with excellence . 298
To function in accordance with ethike arete 299
To function in accordance with the excellences of intellect 300
(6) Personal development 301
(7) Freedom and other "potentialities" 302

CHAPTER SEVEN: "The objective list theory" II: A critical discussion of the different non-internalist pluralist claims 303

Arguments against the "theory" 303

Arguments for the "theory" 307
"Non-subject-oriented" arguments 308
The appeal to the critical power of the theory ... 308

Some "atomist" arguments....309
Griffin's mixture of "definitional" and "radical" deliberation....310
The subject-oriented arguments....313
Subjectivist justification....314
Quasi-subjectivist justification....315
Objectivist justification....317
Perfectionist accounts (justifications)....319
"Non-perfectionist" human nature accounts (justifications)....322

A critical discussion of four different human nature accounts....324
The first Aristotelian appeal to Human Nature: the ergon argument....324
The basic need account....328
(i) Basic needs are teleological needs....328
(ii) Basic human needs are (typically) universal needs....332
(iii) Basic needs are needs which "flow" from human nature....332
(iv) The goals of our basic needs are always valuable....334
(v) The value of basic need satisfaction....336
Criticism of the basic need account....338
The second Aristotelian appeal to Human Nature: On why filia is nonderivatively good for us....340
Griffin's human nature account....345

A general discussion about Human Nature and its evaluative relevance....349

Weaker versions of "the objective list theory"....362

CHAPTER EIGHT: My own mixed theory....367
A short recapitulation of the earlier chapters....368
The theory itself....372

APPENDIX A: How goodness-for differs from, and is related to, other kinds of value 379
Goodness-for vs. subjective value 379
Value-period 380
Intrinsic vs. extrinsic value-period 388
Final vs. instrumental value 390
Agent-neutral and agent-relative value 392
Value-for and agent-relative value 394
Final value-for and final value-period 395

APPENDIX B: Some further characterization of value-for 399
Normative relevance and "action-guidance" 399
The normative relevance of value-for 400
To what extent, and in what ways, value-for-beliefs are action-guiding 403
The different ways in which conceptions of prudential value are action-guiding, or: Why the questions of prudential value is important to us 407
Why we should regard the questions of prudential value as important; normative relevance revisited 410
A comparison with Scanlon's perspectives 411
Is there such a thing as the most plausible answer to the questions of prudential value? 413
The idea that value-for is fundamentally comparative. 414
Value-for is supervenient, but in a very special way ... 416

APPENDIX C: Well-being, value-for, and time 422

APPENDIX D: Subjectivism and Objectivism. Some other important questions of well-being 428

APPENDIX E: Hedonism and time: On the issue of duration 440
Objective duration 441
Two conceptions of subjective duration: the atomistic-summative and the holistic 441
Three different versions of pure hedonism....444
Why it is the holistic kind of subjective duration that matters 445
Does duration (per se) matter at all? 447

APPENDIX F: What conception of desire is most plausible? 452
Phenomenological conceptions of desire 452
Functional conceptions of desire 453
Action as Manifestation of Desire: Desire as a theoretical and explanatory notion 455
Desires and Action-tendencies: Brandt's theory....457
The relation between desire and experience....460

SUMMARY 463

BIBLIOGRAPHY 473

INDEX....481

Chapter One

The questions, the traditional answers, and some notes on "method"

This is an essay about "the good human life" (or well-being), primarily about what it is and what it is not, but I will also touch upon some related issues, e.g., on whether there are such things as universal prudential values, and to what extent (if any) a person's well-being can be measured. The issues that will be discussed are all of a theoretical nature, i.e. there will be no discussion of more "practical issues", like how we should act and think in order to make our lives better.

The main purpose of the book is to find the most plausible answers to the following three substantive questions of prudential value: **(I)** What kinds of situations (facts) have final value (positive or negative) for a person?; **(II)** How do we determine just how good (or bad) a certain situation is for a certain person?; and **(III)** How do we determine how well off a person is (on the whole)?[1]

The way in which I will try to achieve this end is not very original: I will first take a look at how these questions have (in fact) been answered by various thinkers, and when a certain answer is too vague or ambiguous, I will suggest a more precise formulation of it. I will then try to determine how plausible these "theories of prudential value" (or "conceptions of well-being") are, viz. by looking at the reasons that have (or can) be given for and against these theories. This critical discussion of a number of existing theories constitutes the major part of the book, and in the course of this discussion, my own theory will slowly take shape.

But this discussion cannot (and should not) begin until we have a firm grasp on the three questions. It is important that the questions of pru-

[1]Whenever I refer to (I), (II), or (III) in the text, I will have these central questions in mind. (There is one exception to the rule, however, viz. on p 282 ff.).

dential value are formulated in a clear and precise way, and one of the main purposes of chapter 1 is to provide such formulations, mainly by offering a general account of prudential value (or "value-for"). I think there are good reasons for adopting such a "question-oriented" approach: it is, after all, rather obvious that if a question is given a clear enough formulation, if one can get a firm grasp on it, then it is also easier to find (or recognize) a plausible answer.

It is worth pointing out that my three central questions are not the only questions which have figured in the philosophical discussion on well-being. There are many questions of prudential value besides the substantive evaluative questions, e.g., metaethical or methodological questions like "What is the nature of well-being?", "Is it possible to justify (or refute) a substantive conception of prudential value, and if so, how?", and "To what degree (if any) is the prudential value that a certain situation (or life) has for a certain person measurable?". Now, even though most of these questions will be either ignored or discussed only "in passing" (e.g., in appendix D), I think it is important that they are not excluded altogether; partly because they will constitute a fruitful contrast to the three central questions.

Chapter 1 consists of three main sections. In section 1.1, my main purpose is to formulate the central questions of prudential value as clearly and precisely as possible, but there will also be some discussion of why they are important questions, why (and how) they are of normative relevance, how they are related to other questions, and so on. In section 1.2, I offer a survey of the traditional answers that have been given to the central questions. It should be noted that the purpose of this section is solely to give a rough characterization the major theories of prudential value, e.g., there will (at this point) be no discussion of how plausible these theories are. There will then (in section 1.3) be a brief and very general "epistemological" discussion of what general requirements a substantive theory of prudential value must satisfy in order to count as plausible, or more specifically, of how general value-for-claims can be justified or refuted. The main reason for including this discussion is this: In order to determine whether a general substantive value-for-claim is plausible or not, we will (of course) have to assess the arguments which can be given for and against the claim. But how do we do this: how do we determine whether a certain argument is good or bad? What constitutes (in this context) a good argument? Here, I

will simply state my position on this issue: there will be little or no argument for the "general methodological convictions" expressed.

1.1. The questions of prudential value

The main purpose of this section is to formulate the central questions of prudential value as clearly and precisely as possible. In order to achieve this goal, I will (roughly speaking) proceed as follows: I will first offer a general account of prudential value (or "value-for-people"[2]) - of what kind of value it is, to what it can meaningfully be attributed, and so on[3] - and I will then offer some kind of "analysis" (in a broad sense of the term) of the notion of well-being. So, let us first take a closer look at what kind of value value-for-people is, and what implications this has for the understanding and interpretation of our central questions.

Goodness-for-people: What it essentially consists in

Let us start by pointing out that goodness-for-sentences normally take the form "X is (was, will be, or would be) good for Y", or "It is (was, etc.) good for Y that X holds", where Y ("the subject") can be a carpet, a plant, an engine, a company, a person, or the like, and where X ("the object") is "ultimately" an event or a state of affairs. This suggests that goodness-for is really *a relation between an object and a subject*, and if we want to determine whether a certain object is good for a certain subject, we must therefore take into account both what the object is like and what the subject is like.

The only kind of subject that is of interest in this context is people (or human beings), i.e. the only kind of goodness-for that is of any relevance here is goodness-for-people. And again, it is important to notice that goodness-for-people is (like all kinds of goodness-for) ultimately "possessed by" events or states of affairs, and only derivatively possessed by other things (cf. Thomson (1992), p 97). That is, I will restrict my attention to goodness-for-statements of the form "X is (was, will be, or

---

[2]As far as I can tell, this term was coined by Thomson. Cf. Thomson (1992).
[3]This account will be further developed (more than what is necessary to get a firm grasp on the central questions) in appendix A - which deals with how value-for differs from (and is related to) other kinds of values - and in appendix B - where value-for is further characterized.

would be) good for P", where X (the object) is a "situation"[4], and where P (the subject) is a person[5].

Now, to say that "X" refers to a situation, and "P" to a person, is far too ambiguous. If we look at how goodness-for-people-statements are actually used, we see that the terms "X" and "P" may refer to a number of different things. If we look at the term "X", we see that, first, it may either refer to a particular situation or to all particular situations of a certain type; second, the situation X may either be actual (e.g., concrete) or merely possible; and, third, it should also be noted that situations differ "in size", some situations are (so to speak) "larger" (more global) and other situations are "smaller" (more local). In a similar way, the term "P" may refer to different things. First, it is likely that it refers to one or several particular persons (actual or possible), but it may also refer to "generic man"[6]; second, it may either refer to a certain particular (actual) person, or to all (actual) persons, or to all (actual) persons of a certain type; and third, it seems that it may not just refer to one or several actual persons, but also to hypothetical people.

This implies that goodness-for-people-statements may be of many different kinds, depending on what the terms "X" and "P" refer to. In this context, the most interesting kinds of statements are the following ones: (1) "The particular situation X is good for the particular person P", e.g., "it is good for John that he has this particular pleasant experience right now"; (2) "All particular situations of type X (actual or hypothetical) are good for the particular person P", e.g., "it is good for John to have pleasant experiences"; and (3) "All particular situations of type X (actual

---

[4]In this book, I will use the terms "situations" and "facts" as substitutes for the phrase "events and states of affairs". In this terminology, a particular situation (or fact) is (roughly) a concrete part of reality that corresponds to a true statement (sentence, or proposition), and which "makes" this statement true.

[5]At this point, it is worth mentioning that the type of person which I have in mind is a "reasonably healthy", "minimally autonomous", and (on top of this) adult human being (this is something I have in common with most other philosophers of well-being). That is, when I ask what is good for a person, I assume that the person in question is (roughly) a biological organism who is also a conscious, self-conscious, social, and language-using agent. (That is, the terms "persons", "people", and "human beings" can, in this context, be regarded as more or less synonymous). However, this does not mean that a conception of well-being does not have anything to say about what is good for squirrels, infants or gravely handicapped people; it is just that these cases are (with the possible exception of hedonism!) not the central cases in this context.

[6]Or "the generic person"; but cf. note 10 below.

or hypothetical) are good for all persons P" (where hypothetical people may also be included), e.g., "it is good for anyone to have friends"[7].

Even though statements of type (1) are (most probably) most fundamental in the epistemic sense, it seems that the most general type of goodness-for-people-statement, i.e. (3), is most "fundamental" in another sense, viz. because statements of the other two types can be deduced from statements of type (3). This explains why the question "What types of situations are good for everyone?" is often regarded as the central (or fundamental) question of prudential value. But it is worth noticing that this is not generally accepted, and the reason for this is that the question is based on the assumption that there are valid statements of type (3), but this assumption may not be true. If this is the case, i.e. if "goodness-for-relativism" is true, then we have to replace this question with a more "neutral" question, viz. "How (according to what criterion) do we ("in principle", that is) determine what situations are good for a particular person?". This is a perfectly neutral question that everyone can accept, whether they are "generalists" or "particularists", "universalists" or "relativists", "substantivists" or "formalists".

So, what do we mean when we say that a certain situation is good for someone? How should good-for-people-statements (as they are used in ordinary language) be analysed? Well, it seems that sentences of the form "X is good for P" must be analysed in terms of P's interest, or P's well-being, or P's welfare, or P's good, or the good of P, or P's health, or P's being in a good condition. Here are some examples of how the phrase "X is good for Y" can be analysed: According to von Wright (1963), something is good for a being (or beneficial to this being) when the doing or having or happening of this thing affects *the good of that being* favourably (cf. p 45). He also suggests that when the being in question is a human being, the phrase "the good of a being" (or "the being's good") can be understood in two different ways: in terms of *welfare*, and in terms of *health*, where the welfare of a being is the good of the being "as a whole", bodily health is the good (or welfare) of the body, and mental health is the good of the mind. This analysis is almost identical with the one that Thomson (1996) offers: "What is it for a thing X to be good for a thing Y? X's being good for Y presumably

[7]If some kind of cultural relativism is adopted, statements of the type (4) "All particular situations of type X (actual or hypothetical) are good for all persons of type P (e.g., for men, or for traditional people)" may also be of interest.

consists in X's being conducive to Y's welfare, or to Y's being in good condition, or anyway to Y's being in better condition than it would otherwise be" (p 140). If we assume that being healthy and being in a good condition is the same thing[8], we see that goodness-for-people is (again) analysed both in terms of welfare and in terms of health. Two years earlier, Thomson "defined" goodness-for in terms of *"interest"* and *"well-being"*: "What is good for a creature is what is on balance in its interest, or constitutes or contributes to its well-being" (Thomson (1994), p 13). This suggests that when we say (in everyday talk, that is) that something is good for someone, we either mean that it is in some way conducive to his welfare (or well-being), or that it is conducive to his (bodily or mental) health, or both.

This analysis makes it "natural" to make the following two distinctions between different kinds of goodness-for. The first distinction is the one between "health-goodness-for" and "welfare-goodness-for", where something is good for a person in the former way if it conducive to his health, and where something is good for a person in the latter way if it conducive to his welfare (or well-being). In the present context, it is "welfare-goodness-for" that is of interest.

The second distinction (which crosses the first) has to do with how, in what way, the object to which goodness-for is attributed is conducive to a person's health or welfare. The idea is that situations can be conducive to a person's health or welfare in two major ways; by being a constitutive part of it, or by contributing to it in some other way, e.g., by making it possible, or by making it more likely.

If we restrict our attention to welfare, the relevant distinction is the distinction between being conducive to a person's welfare by being a constitutive part of it, and being conducive to a person's welfare by contributing to it in some other way. This distinction is (I think) identical with the distinction that Thomson (1992) made between *derivative and nonderivative goodness-for*. In Thomson's terminology, "the derivatively good inherits its goodness, and the nonderivatively good does not" (p 99). She also points out that a thing can inherit its goodness from another good thing in several different ways, e.g., by causing it, by making

---

[8]It should be noticed, however, that in Thomson's own terminology, being healthy and being in good condition is not the same thing. For Thomson (1996), being in good condition seems to be more or less identical with welfare (or being well off, or having a good life).

it possible, or the like[9]. And this seems to imply that to be nonderivatively good for a person is the same thing as to be a constitutive part of his[10] welfare. There is a problem with this view, however: It seems reasonable to assume that if a situation X is part of another situation Y, and if the value of Y is a function of the values of its parts (including X), then the value of Y can (and should) be regarded as derived from the value of X. But this allows for the possibility that there are constitutive parts of a person's welfare (viz. certain larger wholes) that do not have nonderivative value for this person. This means that we will (in this book) not restrict our attention solely to nonderivative value-for. The focus will be on nonderivative value-for, however (it is, after all, the fundamental thing); it is just that we will also (at times) have the wider category of *final value-for* in mind (viz. in the cases where the distinction between derivative and nonderivative value do not coincide with the more well-known distinction between instrumental value (value as a means) and final value (value as an end)[11], e.g., when we are concerned with the value-for of additive wholes).

All this suggests that the most fundamental question of value-for (or prudential value) can be formulated as follows: How (according to

---

[9]The fact that X inherits its goodness-for-P from Y by making it possible can (I think) be expressed as follows: The statement "X is good for P" can be *derived* (inferred) from the statements "Y is good for P" and "X makes Y possible", or alternatively put: X is good for P *because* [X makes Y possible *and* Y is good for P]. But notice that this "inference" is not of a deductive kind.

[10]His or her, that is. As Jonathan Lear writes in his *Love and its Place in Nature* (in note 1, p 4): "Unless I am specifically referring to a male, "he" should be understood as meaning "he or she", etc. The following observation about the problem of gender pronouns made by David Velleman in *Practical Reflection* strikes me as apt:

"Some readers make take offense at my use of 'he' to denote the arbitrary person. Let me assure these readers that I share their goal of inclusiveness in language and differ with them only about the means to that goal. My view is that traditional usage in this case makes English more inclusive, not less.

"The rule governing traditional usage is that when 'he' denotes the arbitrary person, its gender is purely grammatical, not semantic, and hence carries no implications as to the referent's sex. So understood, 'he' no more denotes a man, because being masculine, than the German 'die Person' or the French 'la personne' denotes a woman, because of being feminine.

"The alternative practices that are currently recommended as inclusive - such as saying 'he or she' or altering 'he' with 'she' - actually threaten to rob the language of its capacity for gender-neutral reference to persons."

[11]The distinction between final and instrumental value is further elaborated in appendix A.

what criteria) do we determine what situations (facts) that are non-derivatively good (and bad) for a particular person? This question may also be formulated in terms of interests, or welfare, or the like. If we do this, we will get: How do we determine what is (ultimately) in a certain person's interest? How do we determine what a person's welfare (or good, or well-being) ultimately consists in?[12] These are perfectly neutral formulations of the question of prudential goodness, formulations that everyone can accept. But how (if at all) could this question be answered?

Well, it is often assumed that there are certain (central, or important) types of situations that are nonderivatively good for everyone, e.g., to feel pleasure, to have one's desires fulfilled, or to have friends. Now, if this assumption is true, the question is really what kinds of situations that are nonderivatively good for everyone, since if we know this, it is easy to determine what situations that are nonderivatively good for a particular person; we simply apply this general knowledge in the particular case. On this "generalist" view, the most fundamental question of prudential goodness can (and should) be formulated as follows: What kinds of situations are nonderivatively good (and bad) for a person? (where "a person" should be interpreted as "all persons" or "any person"). If we formulate this question in terms of interests, or welfare, or the like, we may get: What is ultimately in a person's interest?[13] What has (universal) prudential value?

The reason why I have chosen to regard this substantive evaluative question as the fundamental question of prudential value (cf. (I) on p 1) is simple: I believe that the most central and important prudential values are universal. The only possible threat to this idea comes (as I see it) from a certain kind of subjectivism (or desire theory), but when it is recognized that this view can also be plausibly interpreted as a substantive universalist theory, there is really no need to deny that there are

---

[12]Notice that in this book, I use the term "welfare" is in a somewhat technical sense, a sense which differs from some of its ordinary senses. Cf. the section on the notion of well-being below.

[13]An example of an "analysis" of the question of the good life that seems to identify this question with the question of what is in a person's interest can be found in Parfit (1984). He writes: "What would be best for someone, or would be most in this person's interests, or would make this person's life go, for him, as well as possible? Answers to this question I call *theories about self-interest*" (p 493, Parfit's own italics).

universal prudential values (we will return to the desire theory and its different interpretations below, in section 1.2)[14].

This is (I think) all the background that is needed in order to get a sufficient grasp on our first central question. But before we turn to how the second and the third central questions should be understood, we need to say something about how (I) should *not* be understood; in particular, we need to say something about how value-for differs from "value-period", and how final value-for differs from final "value-period"[15].

Value-for vs. value-period

"Goodness-period" (or absolute goodness, or ethical goodness, or whatever one may like to call it) is the kind of goodness that moral philosophers have paid most attention to, e.g., it is the kind of goodness that is considered most normatively relevant (it is, supposedly, the kind of goodness that utilitarians think we ought to maximize). Moreover, the traditional theories of prudential value, e.g., hedonism or the desire-fulfilment theory, are normally conceived of as theories of "value-period" (or rather: as theories of intrinsic, or final, "value period").

So, what kind of value is value-period, and (most importantly) how does it differ from value-for? This is a very rough answer to this question (a more detailed answer will be given in appendix A): (1) To be good-period is to be good "absolutely", or "to make the world (as a

---

[14]An example of a philosopher who do not agree with me is Sumner (1996). On his view, the fundamental question is not the substantive "What are the (direct, intrinsic) "sources of welfare"?" (i.e. "What is intrinsically good for us?"), but the formal "What is it for something (anything) to be a source of welfare?". This is intimately connected to the idea that a theory of welfare must be formal, i.e. that it must offer us not merely a list of "sources of welfare", but also an account of what qualifies something (anything) to appear on that list (cf. p 16). We will return to this idea several times, but suffice it to say here that Sumner does not give us any reason to believe that the satisfaction interpretation of the desire theory should be rejected.

[15]Goodness-for also has to be kept distinct from goodness-from-a-point-of-view (or subjective value) as well as agent-relative value. These distinctions are discussed in appendix A, which also contains a more detailed discussion of the distinction between final goodness-for and final value-period. In the same appendix, there is also a discussion of how value-for (in general) is related to "value-period" (in general; both agent-relative and agent-neutral value-period), and, in particular, how final value-for is related to final value-period.

whole) a better place" (a better-period place, that is). (2) It is likely that the notion of value-period must somehow be understood in normative terms, e.g., in terms of what we have a reason to promote (aim at), want, or like. And even if value-period-sentences might not (strictly speaking) be *reducible* to normative sentences (in a broad sense), it still seems that we have to regard sentences like "if two situations X and Y can both be realized: X is better-period than Y if and only if we have a prima facie reason to aim at (promote, etc.) X rather than Y" or "X is better-period than Y if and only if it is the case that if some agent could realize X instead of Y, he should do so" as platitudes. In short, it can hardly be denied that there is a strong conceptual connection between "X is better-period than Y", on the one hand, and claims like "If there is a choice between X and Y, we have a prima facie reason to choose X" or "We have a reason to prefer X to Y (on the whole)", on the other.

This strongly suggests that the notion of value-for and the notion of value-period are two entirely different notions. To see this more clearly, consider the following circumstances: (a) It is not impossible that there exist situations which have final value for people, but which lacks final value-period. For example, the alleged fact that it has final value for a person that a certain desire of his is fulfilled does not necessarily make the world better-period, and it does not "imply" that there is an "ultimate" reason to promote that the desire is fulfilled[16]. (b) It is also possible that there are situations that have final value-period without being (finally) valuable for anyone. First, it is (obviously) the case that a situation can have final value-period without also having final value for some existing person: That a certain situation has final value-period (or that a certain situation is finally better-period than another) does not imply that there is any particular (existing) person for whom it is good. To see this, consider the following statements: "It is better that 100 babies are born than that 50 babies are born" and "it is better if 300 people die than if 700 people die". Here, it is quite clear that there need not exist any person (or sentient being) P such that "better" can be replaced by

[16]Or alternatively put, it is not necessarily irrational to accept preferentialism *qua* conception of prudential goodness and (at the same time) reject preferentialism *qua* conception of final (agent-neutral) value-period altogether. (It is worth noticing that this idea is closely related to, but not identical with, the idea that preferentialism is plausible *qua* axiological component of egoism, but that it is implausible *qua* axiological component of utilitarianism).

"better-for-P". Second, and more importantly, the final value-period of a certain situation need not be a function of the final values-for that it "contains". Or alternatively put, all final values-period need not be "personal values", there may also exist "impersonal final values", like distributive equality or ecological diversity.

The fact that the theories of the good life that will be discussed in this book (e.g., hedonism or the desire theory) are often conceived of as theories of final value-period rather than as theories of prudential value makes it important to distinguish carefully between the two types of theories. For example, viewed as a conception of prudential goodness, the hedonistic theory claims (roughly) that the only thing that is ultimately in our interest is to have pleasant experiences. Viewed as a conception of final value-period, on the other hand, hedonism claims (roughly) that the only thing we (ultimately) have a reason to promote is that people's lives (or the lives of sentient beings) are as pleasant as possible. These are two very different claims, and they have to be justified (or refuted) in somewhat different ways.

A note on Betterness-for

In order to make the central questions (II) and (III) more precise, it is necessary to dwell on two rather trivial "claims" about value. The first one is that *evaluations are often comparative*, i.e. they are often of the form "X is better (or worse) than Y"[17]. In the case of value-for-people, comparative sentences either take the form "It is better for P that X holds than it is that Y holds", or they take the form "X is better for P than what Y is for another person Q". The former case is *the intrapersonal case*, and here "every human being" may well be substituted for "P". The latter case is *the interpersonal case*, and here "P" and "Q" either refer to particular persons or to all persons of certain types. In a similar way, we can also distinguish between intratemporal and intertemporal comparisons, where an *intratemporal comparison* has (in the intrapersonal case) the form "it is (at a certain time t) better for P at t if the possible situation X holds at t than if the possible situation Y holds at t", and where an *intertemporal comparison* has (again, in the intrapersonal case)

[17]Comparative evaluations may (of course) also take the ("superlative") form "X (a particular K) is the best (or worst) of all Ks", but evaluations of this type are of little or no interest in this context.

the form "X's occurrence at t_1 is (or was) better for P at t_1 than Y's occurrence at t_2 is (was) for P at t_2" (e.g., as in "my present job is better for me now than the job I had two years ago was for me back then")[18].

The second (trivial) claim is that *value is always a matter of degree.* All good things have (it seems) positive value to some degree or other, and all bad things have negative value to some degree. So, in the case of value-for, we can say that if something is good (or bad) for a person P, then it is always good-for-P (or bad-for-P) to a certain degree. This notion of "value to a certain degree" is (I think) intimately connected to the notion of comparative evaluation, or more precisely, it is likely that "specifications of degrees of value cannot be obtained without recourse to comparative evaluations" (cf. Wetterström (1986), p 85).

Now, these claims are both relevant in this context. The second claim is an assumption that has to made if the central questions (II) and (III) are to make any sense at all, and the fact that there are several different kinds of comparative evaluations implies that there are several ways in which the two questions can be interpreted.

Let us look at (II), which has been formulated as follows: "How do we determine just how good (or bad) a certain situation is for a certain person? For example, how do we compare different possible situations with respect to their final value for a certain person?". This is a question about intrapersonal measurement: what we want to find out is how we should (on the assumption that this is possible) measure to what degree a certain good (bad) situation is good (bad) for a certain person. So, what type of intrapersonal measurement (e.g., comparisons) do I have in mind, intratemporal or intertemporal? Well, the answer is really "both" (it is hard to see how one could suggest different standards of measurement in the two cases[19]), but for reasons that will be given below (e.g., in appendix C), the focus will be on the intratemporal case. That is, (II) will (primarily) be understood as follows: To the extent that it is possible, how (according to what criteria) do we determine just what final value a certain possible situation has for a certain person-at-

[18]These formulations are all "based" on the view that all value-for-P is value-for-P-at-some-time-*t*. This view is formulated in more detail in appendix C.

[19]In a similar spirit, we can add that the standard of measurement that is most plausible in the intrapersonal case is probably most plausible in the *inter*personal case too. But it is far from certain that the two standards coincide fully (e.g., if some desire theory is accepted; cf. note 47 below), and to the extent that they differ, I will ignore the interpersonal question.

a-certain-time? For example, how do we determine whether a certain (possible) situation is nonderivatively better or worse for a certain person-at-a-certain-time than another situation (or whether the two situations are equal in value)?[20] [21]

The notion of well-being

Let us now try to get a better grasp on the third central question, viz. "How do we determine how well off a person is (on the whole)?". In order to understand this question, we need to know what we mean when we say that someone is well off[22], or that his level of well-being is high (or low), or that there has been an increase (or decrease) in well-being. How should such attributions of well-being be interpreted? Well, to say that someone is well off is (obviously) to make an evaluative claim, and the question can therefore be rephrased as follows: When we say about a certain person that he is well off, what kind of value is being attributed, and to what is this value being attributed?

My answer to this question is that when we say about a person P that he is well off, we attribute final (but not nonderivative) value-for to his existence as a whole. Or alternatively put, to be well off is (roughly) the same thing as living a good life (in a broad sense of the term; cf. e.g.,

---

[20]This suggests that (III) is (primarily) a question about how we should (in principle) determine (measure) how well off a certain person-at-a-certain-time is at that time. We will soon return to how (III) should be understood.

[21]The fact that these questions are (in a way) based on the assumption that final value-for is (at least to some degree) intrapersonally measurable gives rise to a number of corresponding questions concerning *measurability* (e.g., *comparability*). Examples of such questions are: Is the final value that a certain (possible) situation has for a certain person intratemporally measurable, and if so, in how strong a sense? For example, is it possible to compare (rank) all possible situations with respect to their final value for a certain person-at-a-certain-time? And in how strong a sense (if any) is well-being (intrapersonally) measurable, e.g., is it possible to rank all possible lives for P at *t* with respect to value for P at *t*?

Now, it is worth noting that even though it might be of great normative importance to find plausible answers to these questions, they are not questions that a conception of prudential value is supposed to answer. Every theory of well-being has certain methodological implications, however, and to find out what these implications are, all we have to do is to assume that the theory is true, and then ask in how strong a sense (if any) that well-being is measurable (given this assumption). These methodological questions will fall outside the scope of this investigation, however.

[22]What we mean in this context, that is, and not in ordinary speech. I will, for example, ignore the fact that "being well off" sometimes means "being financially well provided for".

note 49 on pp 224-225)[23]. But what is a person's existence (or "life"), and what is it to attribute value-for to such a thing?

The situations (facts) that have final value for a person differ in "size" as well as in complexity. That is, the situations to which final value-for is attributed can be more or less local (more or less global), and they can be more or less simple (more or less complex). The smallest (most local) situations to which final value (for a person P) can be meaningfully attributed is (I think) situations like "P has such-and-such an experience (e.g., a pleasant experience) right now", "P is engaged in such-and-such an activity (or performs such-and-such an action)", or "P interacts with Q in such-and-such a way (e.g., he has sex with Q)". Examples of possible prudential values with a higher degree of globality are accomplishment, intimate relationships (where a relationship should be regarded as a much more global and complex thing than an interaction), personal development, and autonomous living. Now, my idea is (roughly) that the most global (and the most complex) situations to which final value-for-P can be meaningfully (and plausibly) attributed are situations of the type "P has such-and-such an existence"[24], where a P's existence (or "life" in the broad sense) includes such things as what kind of person P is, and what kind of life (in the narrow sense) he has[25].

The idea that "P is well off" can be understood as "P's existence (as a whole) has positive final value for P" suggests that the central question (III) may also be formulated as follows: How do we determine just how

---

[23]That is, the fact that well-being is attributed to persons (we do not say about a person's existence that it is well off) does not in any way imply that we attribute value to a person when we say about him that he is well off. And once this is clear, we will hardly assume (with Kagan (1992)) that "it might be one thing for a person to be well off and quite another for that person's life to go well" (note 7, p 182), or that the quality of a person's life may be distinct from his level of well-being (cf. ibid., note 10, p 188).

[24]It is worth noticing that this idea is inconsistent with the unrestricted version of the desire theory. The reason for this is that this theory attributes final value-for-P to the most global situations which can be imagined, viz. situations of the type "the world (as a whole) is constituted in such-and-such a way". As far as I can see, the idea is consistent with all the other theories, however, e.g., the fact that the hedonist attributes final value to our experiential lives only does not seem to prevent him from attributing final value-for-P to a P's existence as a whole (it is just that this value is believed to "reside" solely in the experiential dimension).

[25]We could also say that a full description of a person P's existence would take the form of a list of factual claims about P and his life. But what types of facts should be included in such a description, and what facts should be excluded? We will return to this issue in section 5.2.1.

finally valuable a person's existence is for this person? Or in terms of information: What do we need to know about a person's existence in order to be able to determine its final value for this person? For example, how do we determine whether a certain possible life is better or worse for a person than another possible life, or whether the two lives are of equal value?

Now, it seems that "P's life (existence) is good for P" is better formulated as "P's life is a good life for P" (or as "P's existence is a good existence for P"). Here, the term "good" is affixed to the noun "life", i.e. it is used adjunctively. This suggests that when we say that a certain life is good for the person who lives it, we attribute goodness-for to a certain kind of entity, viz. a life (or an existence). It also suggests that "good-for" is (in this context) an attributive adjective in Geach's technical sense, i.e. that statements of the form "X is a good life for P" never split up logically into "X is a life" and "X is good for P" (cf. appendix A). Or alternatively put: The goodness-for that we attribute to existences belongs to the category "goodness-of-a-kind"[26]. This explains why the non-comparative question of well-being (which is really a part of (I), and which should be kept separate from the comparative (III)), viz. "What does a person's well-being consist in?" can also be formulated as "What *kind* of life is a good life for the person who lives it"?

Let us now connect this to the idea of supervenience. In the case of attributive goodness, the idea of supervenience can (I think) be formulated as follows (cf. appendix B): For each kind K, there is a standard of goodness that is of the form "a good K is a K that has the natural features $F_1 \ldots F_n$, and if X is a good K, it is because it has these features". That is, as far as attributive goodness is concerned, a standard of goodness is always a standard of goodness for some kind K, and for each K, this standard is common to all Ks. This suggests that the non-comparative question of well-being can also be formulated as "What is it that makes an existence (a life) good for the person who has (lives) it?", or "What are the "good-for-making characteristics" for human existences (lives)?".

Now, in order to get a better grasp on how the questions of well-being (especially (III)) should be understood, there are at least two

---

[26]That is, value-for is sometimes attributive, and value-for-statements are sometimes of the form "X is a good-K-for-P". "K" need not refer to existences, however; it can also refer to experiences, activities, relationships, or the like.

more things we have to know. First, phrases like "P's existence" and "P's life" can either refer to P's existence (life) at a certain time, or to P's existence (life) over time, i.e. to something temporally extended. So the question arises: How should the phrases be understood in the present context of well-being? Second, regardless of whether we have existences at certain times or temporally extended lives in mind; existences (lives) are wholes, and this has certain implications for how the questions of well-being should be understood and answered, implications which need to be spelled out. Let us now take a closer look at these issues.

Well-being and time: the temporal aspect of (III)

So, depending on whether we have existences at certain times or temporally extended lives in mind, there are two possible interpretations of the questions of well-being. The non-comparative question of well-being (i.e. "What makes a life have final value for the person who lives it?") can either be interpreted as (i) "What makes a person's life good (for this person) at a certain given time[27]?", or as (ii) "What makes a person's life good (for this person) over a (longer) period of time?" (a special case of which is (iii) "What makes a life as a whole (from birth to death) good for the person who lives it?"). The comparative question of well-being (i.e. (III) "How do we determine just how valuable a person's existence is for this person?") can, in a similar way, either be interpreted as (i) "How do we determine just how well off a certain person is at a certain time?", or as (ii) "How do we determine just how valuable a person's existence (life) over a certain period of time is for this person?" (a special case of which is (iii) "How do we determine just how valuable a certain life as a whole (from birth to death) is for the

[27]So, how should the phrase "at a certain given time" be understood here, how extended is the time span that it refers to? Examples of "certain given times" are "now" and "then", as in "my life is good now, but it was not good then (a year ago)". So, when someone says that his life is good now, what time span does he have in mind? Well, "now" does most probably not mean "this very moment", but rather something like "nowadays". So, what does one mean when one says that one's life is good nowadays? Presumably that it has been good for some time (at least for a couple of days), but also that it is likely to stay that way (at least for some time). If it is true that judgements of the form "P's life is good right now" involve an implicit reference to the near future (or at least to beliefs about the future) in this way, then every plausible answer to (i) must take this into account, but I have never really seen this done (in a philosophical context, that is).

person who lives it?")[28].

Suffice it to say that in this book, the two questions of well-being will be understood in the first (synchronic) sense, rather than in any of the diachronic senses. That is, the non-comparative "What makes a life good for the person who is living it?" will be understood as "What makes a person P's life at a certain given time have final value for P at that time?", and the comparative (III) will be understood as "How do we determine just how well off a certain person is (on the whole) at a certain time?"[29].

Lives are wholes

Let us now turn to the idea that existences are wholes, and see how this idea might affect our understanding of the questions of well-being. It is (especially in this context) natural to regard lives as *wholes*: when we attribute value-for-P to P's life, we attribute value-for to a certain type of whole. There are two ways in which lives can be regarded as wholes, viz. "in the diachronic" and "in the synchronic". To say that lives are wholes *"in the diachronic"* is to say that they are (can be regarded as) temporal wholes which consist of a sequence (or succession) of parts (e.g., actions and events), and to say that lives are wholes *"in the synchronic"* is to say that they can (and should) be regarded as "simultaneous wholes" that consist of a number of parallel (simultaneously occurring) states, actions, or events. This means that a life-over-time (e.g., a life-as-a-whole) is a whole in both senses, i.e. in the diachronic as well as in the synchronic, while a life-at-a-certain-time-*t* is (roughly

---

[28]It is important to distinguish the idea that the questions of the good life (or existence) has "temporal content" (that one can, so to speak, have different "time spans" in mind when one asks them) from the observation that lives (or periods of lives) are always evaluated from temporal perspective or other, e.g., while they are going on, afterwards, and so on.

[29]My main reason for adopting this "synchronic interpretation" is that I tend to regard it as the only intelligible (and plausible) interpretation. Or more specifically, I think it makes perfectly good sense to talk about (i) the final value-of-a-life-at-a-certain-time-for-the-person-who-is-living-it, but I don't think it makes any sense to talk about (ii) the final value-of-a-life-over-time-for-the-person-who-is-living-it, and/or (iii) the final value-of-a-life-as-a-whole-for-the-person-who-is-living-it. However, the "fact" that it makes little or no sense to attribute final value-for to lives as temporally extended wholes does not (in any way) suggest that it is unintelligible to attribute aesthetic value or value-period to whole lives. These (somewhat controversial) ideas are developed more in detail in appendix C.

speaking) a synchronic whole, but not a diachronic whole. And since we have restricted our attention to what it is that makes a person's life good (for him) *at a certain given time*, this means that we will (from now on) conceive of lives as synchronic wholes.

Now, suppose we ask what kinds of features of lives-at-certain-times that make them good for the persons who live these lives. If we think of lives as synchronic wholes which consist of parts, two kinds of possible answers present themselves. First, a life can (of course) be good because its constituent parts (or "contents") are good, e.g., because there is a lot of pleasure "in" it, or because it "contains" a lot of friendship. But second, it is also possible that the goodness of a life supervenes on certain *"holistic" features*, i.e. properties that are (roughly speaking) "possessed" by the whole but not by its parts. The holistic features of a life-at-a-certain-time that are most likely to be good-making are its *"formal features"*, i.e. features which concern how its contents are structured (synchronically), e.g., organic unity[30].

An example of a philosopher who thinks that organic unity matters, or more specifically, that a more organically unified life is *ceteris paribus* better than a less organically unified life[31], is Nozick (1989). He writes:

In wanting ourselves to be of value and our lives and activities to

---

[30]A life-over-time has more global features, however, e.g., features which concern how the contents of a life are distributed over time (its "narrative direction"), and a life-as-a-whole has (on top of this) also a *duration*. Of all the features of lives, this is the quantitative feature *par excellence*, and this is why a question like "what is the most optimal life span?" can be regarded as a question of *"the quantity of life"*.

But observe that since the first "*t*" in "X-at-*t* is good for P-at-*t*" can refer to a (short) period of time (cf. note 27), it may well be possible to attribute properties like rhythm and continuity to lives-at-certain-times as well.

[31]Where *organic unity* (or *"unity in variety"*) is a combination of the features unity and variety (or diversity). This means that the degree to which something is organically unified is a function of two things, viz. "the degree of [the] diversity [that gets unified] and the degree of unity to which that diversity is brought" (Nozick (1989), p 164). That a life is organically unified *in the synchronic* means that it is both varied and unified at a certain given time, i.e. that it has many parts (or "dimensions"), and that these parts "hang together", or form a coherent and harmonious whole (e.g., the life in question can not be characterized as a "double life", or as a set of "multiple lives"). (Cf. the idea that the good life is a balanced or harmonious life, e.g., by displaying "hierarchy"). That a life is organically unified in the *diachronic* sense, on the other hand, means (roughly) that it is both varied and unified over time, e.g., that there is a rhythm of alternation between things like novelty and adventure, on the one hand, and continuity and tradition, on the other.

> have value, we want these to exhibit a high degree of organic unity. /.../ We want to encompass a diversity of traits and phenomena, uniting these through many cross-connections in a tightly integrated way, feeding these productively into our activities (p 165)[32].

Now, I am far from certain that this idea is a plausible idea about what is good for a person, but I do think it is worth taking seriously. In other words, we can not automatically assume that all the situations which have nonderivative value for a person are "local"; we also have to consider the possibility that it has final (and nonderivative) value for P-at-*t* that his life has certain formal features at *t*, e.g., that it is organically unified at *t*. That is, we should not ignore the question of whether the value (or "quality") of a life-at-a-certain-time is dependent on what "form" it has[33].

Evaluative atomism vs. evaluative holism

Another thing that becomes clear if we think of lives-at-certain-times as synchronic wholes is the following problem: We attribute final value to lives, but also to their features (e.g., parts). But what connection (if any) is there between the value of the whole and the values of its features (parts)? Well, I don't think anyone would deny that there is some kind of correlation between the value of a life-at-a-certain-time (as a whole) and the values of its features (parts), but there are different views as to the nature of this connection. As a general evaluative view, *evaluative atomism* claims that the value of a whole is a function (e.g., a sum) of the values of its parts. The *evaluative holist* rejects this idea, and if he is an extreme holist, he might even claim that a certain feature can

---

[32]This is but a special case of a more general view, viz. the idea that "[s]something has intrinsic value /.../ to the degree that it is organically unified". (He even claims that a thing's organic unity *is* its intrinsic value, or "at any rate", that "it is a structure of organic unity that constitutes value's structure" (p 164)). It is not entirely clear what kind of value he has in mind here, though.

[33]And if it is the value of a life-over-time, or the value of a life-as-a-whole that we happen to be interested in (a value that cannot, on my view, be value-for), we also have to ask: (1) Is the value of a life-as-a-whole dependent on how long it is?, and (2) Is the value of a life-over-time dependent on what "diachronic form" it has, e.g., (a) on how organically unified it is over time, or (b) on how its contents are distributed over time? (Where the answer to (1) is obviously "yes" if one has value-period or "perfectionist value" in mind, but not necessarily if one has aesthetic value in mind, where the answer to (2a) is obviously "yes" if one has aesthetic value in mind, and so on).

be a merit in some lives and a defect in other lives. If we apply this to the case at hand, we can say that someone is a *"synchronic atomist"* if he claims that the value of a life-at-a-certain-time is a function of the values of its "simultaneous" (or "parallel") features/parts[34], and that someone is a "synchronic holist" if he rejects this view[35].

So the question arises: Which view about the connection between the final value-for of a life-at-a-certain-time (as a whole) and the final values-for of its features (parts) is correct, synchronic atomism or synchronic holism? Well, personally, I think that it is quite clear that "synchronic atomism" is false (at least as long as we do not conceive of holistic features as parts), i.e. I agree with Moore's (1903) idea that we are sometimes "justified in asserting that it is far more desirable that a certain thing should exist under some circumstances than under others; namely when other things will exist in such relations to it as to form a more valuable whole" (p 30)[36]. The holism I accept is not very extreme, however. For example, I tend to agree with Russell (1930) when he claims that "[t]here can be no value in the whole unless there is value in the parts" (p 25), and I do not agree with Moore (1903) when he claims "that a good thing may exist in such a relation to another good thing that the value of the whole thus formed is *immensely greater* than the sum of the values of the two good things", or when he says that "it seems as if indifferent things may /.../ be the *sole constituents* of a whole which has great value, either positive or negative" (pp 27-28, my italics)[37]. On my view, the value-for of a life-at-a-certain-time is

---

[34]That someone is a *"diachronic atomist"* would then mean that he believes that the value of a life *over time* is a function of the values of its temporal parts (or its "periods"). However, the value referred to here can (on my view) not be value-for (cf. appendix C).

[35]As it stands, it is not entirely clear how this should be understood, however. In particular, it is not clear whether we should think of the holistic features which were discussed above as "parts". If there are such things as holistic good-making features of lives, does this imply that holism is true and atomism false? Well, I think this is the common view, but I don't think that it is necessarily the most plausible view. There are two separate questions here, and there are good reasons for formulating them in such a way so as to make this circumstance as clear as possible.

[36]At this point, someone might think that we could make things clearer by introducing a notion of *contributory value*. It is doubtful, however, whether we gain anything by doing this. If atomism is true, we don't need it, and if holism is true, it seems that the notion in question does not really help us to clarify anything.

[37]As far as the value-for of a life-at-a-certain-time is concerned, that is. It is worth

"roughly" a function of the values-for of its parts, and this means that I also reject the extreme idea that the same feature can be a merit in one life and a defect in another (under all descriptions, that is)[38]. The form of holism I accept is compatible with "generalism" in this area, i.e. it does not force me to become a particularist[39].

The problem of (synchronic) aggregation

The last problem that is generated by the idea that lives-at-certain-times are wholes is the *problem of aggregation*. Suppose that synchronic atomism is correct, and that the value-for of the whole is a function of the values-for of its parts[40]. This gives rise to the following problem: What is the *function* that takes us from the values-for of the ("simultaneous") parts to the value-for of the (synchronic) whole? How do we calculate the final value-for of a life-at-a-certain-given-time from the final (nonderivative) values-for that it contains at that time? This is *the problem of aggregation at a certain time*, or "the problem of aggregation in the synchronic"[41].

Now, it is (I think) rather obvious that there is at least "some truth" in synchronic atomism: it can hardly be denied that the final value-for of a life at a certain given time is at least partly dependent on the values-

---

noting that both Moore and Russell had value-period (rather than value-for) in mind. There are exceptions to the rule, however: First, the complex fact that P wants X *and* X obtains may well have value even if the fact that P wants X and the fact that X obtains are both (in themselves) "indifferent", and second, the rule does not seem plausible if one has very local (e.g., "microscopic") parts in mind: there is hardly any value in such parts.

[38]But only on the assumption that the satisfaction interpretation (rather than the object interpretation) of the desire theory is adopted. See section 1.2.

[39]But it is worth noting that the idea that the occurrence of a certain situation may be better for a person under some circumstances than under others means that the following questions can not really be avoided (in connection with (II) and (III), that is): "For every type of good situation, under what circumstances is it best for a person that it occurs? What combinations of good-for situations tend to form synchronic wholes that are more valuable for a person than the sum of the values-for of their constituent parts?" These are extremely difficult questions, however, and they will therefore be put aside.

[40]Most answers to (III), the comparative question of well-being, e.g., the hedonistic answer, seem to be based on this assumption.

[41]There is (of course) also a *problem of aggregation over time*, or "a problem of aggregation in the diachronic", viz. how to calculate the value of a life over a certain period of time (e.g., the value of a life as a whole) from the nonderivative values that it contains during that period. This is not a problem that needs to concern us here, however.

for of its constituent parts at that time. This suggests that no answer to the comparative question of well-being (i.e. (III)) can be complete unless it includes a theory of aggregation in the synchronic, and that no analysis of this question can be complete unless it includes a reference to the problem of aggregation.

This ends the section on how the conception of life-at-a-certain-time as a synchronic whole influences (or ought to influence) our understanding of the questions of prudential value, especially the questions of well-being.

The questions of prudential value revisited

We now have the background we need in order to get a good enough grasp on our three central questions (or sets of questions). To sum things up, this is how these questions can be formulated more in detail:

> **(I)** On the assumption that the central and important prudential values are universal: What kinds of situations (facts) have final value (positive or negative) for a person? What does a person's well-being (or welfare) consist in? Or more specifically, what is it that makes a person P's "existence" at a certain given time have final value for P at that time? What features of a life-at-a-certain-time make (if present) this life better (for the person who is living it) than it would otherwise have been?
>
> In particular, we want to know what kinds of local situations that are nonderivatively good (bad) for a person, but we also have to consider the possibility that it has final (and nonderivative) value for P-at-*t* that his life has certain formal features at *t*, e.g., that it is organically unified at *t*[42].

The most general comparative question can be formulated as follows: To the extent that it is possible, how do we measure (intrapersonally as well as interpersonally, and intratemporally as well as intertemporally) the final value that a certain possible situation (local or global) has for a

---

[42]It might also be argued that a person's well-being at a certain time is (in part) a function of what his future prospects are at that time (which would include how long it is expected to be). I don't think that the value that it has for me-now to have certain future prospects (e.g., future possibilities) is *final*, however, and I therefore exclude the question of good future prospects from the investigation (in spite of the fact that good prospects may well have "contributive" value).

certain person-at-a-certain-time? For example, how do we (in the intrapersonal case) compare different possible situations (local or global) with respect to their final value for a certain person? Can we assume that the more there is of a good thing, the better, and can we assume that the value of a certain good is proportional to the amount of this good?

If we restrict our attention to the intrapersonal case, and if we make a distinction between the global case and the more local cases, we get:

(II) In the case of more local situations: To the extent that prudential value is (to some extent) intrapersonally measurable, how do we determine just how (nonderivatively) valuable a certain local situation is for a certain person at a certain time? For example, how do we (in the intratemporal case) compare different (possible) local situations with respect to their nonderivative value for a certain person at a certain time?

(III) In the global case, i.e. in the case of whole existences-at-certain-times: To the extent that well-being is intrapersonally measurable, how do we determine just how well off a certain person is (on the whole) at a certain time, and how do we determine how valuable a certain possible existence-at-a-certain-time would be for the person for whom it is possible? For example, according to what criteria do we (in the intratemporal case) compare different possible "lives-at-certain-times" with respect to their final value for a certain person?[43]

If we realize that there is some truth in "synchronic atomism", we can see that an answer to this question cannot be complete unless it includes an answer to the problem of synchronic aggregation, viz.: To the extent that it is possible, how do we calculate the final value-for of a life at a certain given time from the nonderivative values that it "contains" at that time? Or alternatively, if we assume that the value

---

[43]If we switch to the intertemporal case, we can see how the question how we should determine *changes* of well-being enters the picture. The reason why this question is (so to speak) identical with the question of how to make intertemporal intrapersonal comparisons of well-being is simple: A certain change in a person's circumstances constitutes an increase (or decrease) in well-being if and only if his is better (or worse) off after the change than he was before the change. We do not have to determine a person's *level* of well-being in order to determine whether a certain change is for the better or for the worse, however, nor in order to determine how big a certain increment (or decrement) in well-being is.

it has for P that he has a certain kind of existence is (to some extent) a function of the prudential values of a number of smaller-scale facts about P: How do we calculate the final value that P's existence at *t* has for P from the nonderivative values of the facts that hold of P (at *t*)?

These are the central questions of this essay[44]. But before we look at the traditional answers that have been given to these questions, there is one more issue which needs to be addressed, viz. the following one: It can hardly be doubted that the notion of value-for is a normatively relevant notion, and it seems (moreover) plausible to assume that there are several different ways in which a conception of value-for can be normatively relevant[45]. So the question arises: Does this have any implications for how the three central questions should be understood (and answered)? For example, is it necessary to regard the questions of well-being as relative to some normative context or other, or as asked from some normative point of view or other? And if this is so: is the meaning of such a question something which varies with the normative point of view from which the question is being asked?

---

[44]It is, however, worth noting that these are not the only questions to which a complete conception of well-being must provide answers. In order to see this, we just have to take a look at the recent philosophical discussion of well-being: at what this discussion has been about, and what questions it has centred on. If we do this, we can see that there is (of course) a considerable overlap between the questions that have been discussed in the literature and the central questions of this book, but we can also see that the two sets of questions do not fully coincide. First, some of the questions listed above are often ignored in the literature (especially the questions that are actualized by the idea that lives are wholes, e.g., the question of good synchronic form), and second, some of the central questions in the literature are not included in the central questions of this book (they will be of some importance in this book too, however, but only indirectly). The most important of these questions is the question (or questions) on which "subjectivists" and "objectivists" disagree; questions which will be made explicit (or "reconstructed") in appendix D.

[45]This does not mean that the notion of value-for is a normative (or quasi-normative) notion, however, or that it can (or should) be understood in normative terms. Value-for-statements are "evaluatives" rather than "directives", and *qua* evaluatives, they do not in themselves tell us what we have reason to promote: to do this, they must be combined with some directive or other. This idea is further elaborated in appendix B.

A note on Normative relevance

There are several ways in which the questions of prudential value (and their respective answers) are of normative relevance. Or alternatively put, there are several plausible norms (prima facie or not) which involve a reference to people's welfare. The most important examples of such norms (or "objective reasons for acting") are (i) self-interested norms, like "everyone has a prima facie reason to promote his or her own welfare", and (ii) benevolent norms, e.g., "we ought to promote other people's welfare, especially the welfare of our own children"[46].

Now, this gives rise to the possibility that there is no such thing as *the* most plausible conception of prudential value, i.e. that the most plausible answer to the questions of prudential value may (instead) vary with the different ways in which a conception of well-being may be normatively relevant.

So, is there such a thing as *the* most plausible answer to the questions of prudential value? For example, is the welfare that we have a reason to promote *qua* self-interested *the same human welfare* that we have a reason to promote *qua* benevolent third parties? And are these "welfares" the same human welfare that the interest-utilitarian thinks we ought to maximize?

Well, on my view, the answer is "yes" (but cf. note 47), and this implies that the question (I) "What has final value for a person?" need (roughly speaking) *not* be understood in relation to any normative theory (or *prima facie* norm) N, i.e. it need not be interpreted as "What conception of prudential value makes (if embedded in N) N most plausible?". This also means that it is (as a rule) not really necessary to bring in any normative intuitions in order to determine whether a certain conception of well-being is plausible: straight-forwardly evaluative intu-

---

[46]These are not the only ways in which conceptions of well-being can be of normative relevance, however. For example, such conceptions also supply certain normative theories with the axiology they need, viz. theories that are (in a wide sense of the word) "teleological", i.e. theories that claim that moral rightness or practical rationality consist solely in the maximization of (and/or some distribution of) nonderivative (or final) goodness-for, e.g., like the self-interest theory or "interest-utilitarianism". Moreover, it is also possible that there are valid norms of the malevolent type, e.g., the retributivist idea that we sometimes have a good moral reason to harm other people. And so on. There is a more detailed discussion of all this in appendix B.

itions will do, i.e. we can safely remain in the evaluative sphere[47].

This is all I have to say on the three central questions of this book, and on how they should (and should not) be understood. It is now time to take a brief look at how the questions have been answered.

1.2. A brief overview of the main answers to (I)-(III)

Substantive vs. formal theories of prudential value

The central questions of this essay are all substantive evaluative questions, and the answers to these questions will (therefore) take the form of substantive evaluative claims, e.g., (in the case of (I)) straightforward claims about what is nonderivatively good and bad for us.

Now, even though the focus will (in this essay) be on substantive theories of prudential value (or what Scanlon calls "substantive good theories"), it is important to point out that a theory of prudential value need not be substantive: it may also be *formal*. A formal account of prudential value does not make any substantive claims about what is good and bad for us; it specifies (instead) some *formal criterion* according to which it can be determined what is good and bad for a person. For example, rather than telling us what a life must "contain" (or "consist of") in order to be good, it may tell us that the good life is the kind of life that a well-functioning (e.g., rational, autonomous, or authentic) person would live under acceptable circumstances, i.e. that a life L is a good life for a person P if and only if it is the kind of life that P would lead if he were a well-functioning person living under acceptable circumstances[48].

---

[47]Or more specifically, I agree with Sumner (1996) that a theory of welfare must be both "descriptively" and "normatively adequate", and I also tend to believe that the normative adequacy of a certain theory does not vary from one normative context to another. There is one possible exception to the latter idea, however; the version of the actual desire theory (cf. section 1.2) that is most plausible in an intrapersonal normative context may not fully coincide with the version that is most plausible in an interpersonal context (cf. section 5.1.2). Personally, I don't regard this as problematic, however, since I have already made it clear that I will (in this book) focus solely on the intrapersonal case.

[48]This idea should be carefully distinguished from another formal idea, viz. the idea that a life L is a good life for P if and only if P would *prefer to lead L* if he were a well-functioning person.

In Griffin's (1986) terminology, a formal account of the good life is an "account of the modes of approach that will fix on the [good] life" (p 63). This seems to mean that an account of prudential value is formal only if it specifies some "procedure" or "method" that (if followed) will generate correct substantive views on what is nonderivatively good and bad for different persons. This suggests that an account that is formal in my sense is not necessarily formal in Griffin's sense (an example of this would be the object interpretation of the desire theory, i.e. the idea that X is good for P if and only if P has an intrinsic desire that X holds). But it is (of course) possible that an account that is formal in my sense is also formal in Griffin's sense (an example of this would be the idea that X is good for P if and only if P believes that X is good for him *and* if his belief has been generated by a certain procedure)[49].

To get a better grasp on the distinction between substantive and formal theories, let us now consider the distinction between two possible interpretations of the desire theory, viz. *the satisfaction interpretation* (which is a substantive theory) and *the object interpretation* (which is a formal theory).

The desire theory claims (roughly) that the only thing that has nonderivative value for a person P is that his intrinsic desires are fulfilled, where P's desire that a situation X obtains is fulfilled if and only if X holds. But even though there is (on condition that P desires that X, that is) a very intimate connection between the fact that X holds and the (relational) fact that P's desire is fulfilled, these are still different facts. That is, it seems that we have to distinguish between the circumstance that P's desire is satisfied and the object of the desire (cf. Rabinowicz and Österberg (1996), p 2).

This makes it possible that one of the facts is nonderivatively good for

[49]Here are some examples of formal theories in other parts of ethics: (i) Nozick's idea that a distribution is just if and only if (and because) it has a "fair history"; (ii) the Kantian idea that an action is morally right if and only if it is in accordance with some valid norm, where a norm is valid if and only if we can consistently will that it should become a universal law; (iii) discourse ethics (on one possible interpretation), viz. the idea that norm is valid if and only if (and because?) there would be rational consensus about its validity, i.e. if people would agree about its validity after having been engaged in a certain kind of (rational) discourse; (iv) Elster's idea that a desire (or belief) is rational if and only if (and because) it has been generated in the right way (and so on); and (v) Brandt's idea that a desire is rational if and only if (and because?) it would survive (or be generated by) a process of ideal deliberation.

P but not the other: it may be nonderivatively good for P that his desire that X is fulfilled, but not that X holds, and vice versa. So the question arises: When a desire theorist claims that it is nonderivatively good for a person to have his desires fulfilled, to what (exactly) does he assign nonderivative value-for? "Is it (i) the circumstance *that* our (intrinsic) desires and preferences be satisfied, or is it rather (ii) those states that are the *objects* of our (intrinsic) preferences and desires?" (ibid., p 2). To believe that (i) is the right interpretation of the desire theory is to adopt what Rabinowicz & Österberg calls the *satisfaction interpretation* of the theory (this is what Österberg argues for), and to believe that (ii) is the case is to adopt what they call the *object interpretation* (this is what Rabinowicz argues for)(cf. ibid., pp 2-3).

That is, the satisfaction interpretation of the theory makes a straightforward claim about what has nonderivative value for a person, and it must (therefore) be regarded as a substantive theory. The object interpretation of the desire theory, on the other hand, claims that a situation X is nonderivatively good for a person P if and only if (and because) P has an intrinsic desire that X holds. This is a formal criterion of value-for rather than a substantive claim about what has value for a person, and the object interpretation of the theory must (therefore) be regarded as a formal theory[50].

Now, even though there will be some discussion of different formal theories in this book (e.g., the object interpretation of the desire theory

---

[50]Does this mean that the object interpretation must be regarded as a subjectivist (or "projectivist") theory about the nature of value-for-P, i.e. must it be understood as claiming that the fact that something (X) is good for P consists in (is constituted by, or identical with) the fact that P desires that X? Or does the object interpretation merely presuppose such a theory of the nature of value (as Rabinowicz seems to suggest; cf. ibid., p 19)? In any case, it is rather obvious that some formal theories (in my sense) are not merely answers to "how do we determine what is nonderivatively good for a person?": they can also be (at least in part) regarded as theories about the nature of value, as conceptions of justification, or the like. This is (typically) another feature that distinguishes them from substantive theories (cf. appendix D).

It is worth pointing out that the two interpretations of the desire theory are equally informative (they give the same answer to questions like "what do we need to know in order to tell how well off a person is?"). This shows that substantive theories are not necessarily more informative (or more specific) than formal theories. As a rule, substantive theories are more informative than formal theories, however, e.g., the substantive claim that it is good for all human beings to have friends is (somehow) "more specific" (and more informative) than the formal idea that it is good for every human being to realize his or her potential.

and the basic need account), it is (as we have seen) the substantive questions (I)-(III) which are most central. So let us therefore give a short characterization of the most important substantive theories of prudential value.

The perhaps most common classification of (general) theories of prudential value is the one constructed by Parfit (1984), viz. the distinction between hedonistic theories, desire-fulfilment theories, and objective list theories. Of these three theories, it is only hedonism that is clearly substantive, i.e. it is doubtful whether the last two theories should really be understood as substantive (cf. appendix D). It is possible, however, to interpret both of them as substantive theories (or as having substantive evaluative content), viz. if we accept the satisfaction interpretation of the desire theory, and if we conceive of the objective list theory as a theory that is both "non-internalist" and pluralist[51].

Let us now look at how the three types of substantive theories can be characterized. We will start with the most simple theory, viz. hedonism.

The Hedonistic Theory

The hedonistic theory of prudential value (or "pure hedonism") can be characterized as follows. First, this is how the theory answers **(I)** "What kinds of local situations have nonderivative value for a person?":

> **(H1)** Nothing but a person's own *experiences* (concrete conscious mental states and events, like sensations, emotions, perceptions, moods, fantasies, and thoughts) can be nonderivatively good or bad for this person. Or more precisely, the only local situations (or "atomic facts") that can have nonderivative value for a person P are situations of the type "P has a certain experience E". This feature of the hedonistic theory is sometimes called *the Experience Requirement*. If this require-

[51]What this means will be explained later, but it should be pointed out that the objective list theory is (*qua* substantive theory) not merely a theory that is neither a hedonist theory nor a desire theory. This means that the present classification is not really complete: there are possible views which "fall outside" the classification, e.g., certain monistic theories (like the idea that nothing but love has nonderivative value-for). But if we restrict our attention to theories which have actually been held by various philosophers, the classification is (I think) complete, i.e. it seems that the only serious alternatives to hedonism and the desire theory are both "non-internalist" and pluralist.

ment is included in a theory of well-being, we get the idea that a person's level of well-being at a certain time is a function of one thing only, viz. what his "experiential life" (or total mental state) is like at that time.

(H2) More specifically, the only thing that is nonderivatively good for a person is to feel pleasure (to have pleasant experiences), and the only thing that is nonderivatively bad for a person is to suffer (to have unpleasant experiences). This implies that every experience which is nonderivatively good for a person is also pleasant, and that every experience which is nonderivatively bad for a person is also unpleasant.

(H3) All pleasant experiences are nonderivatively good for the experiencing subject, and all unpleasant experiences are nonderivatively bad, regardless of what other properties these experiences have. That is, the class of pleasant experiences which it is good for a person to have, and the class of unpleasant experiences which it is bad for a person to have, these classes are, on the theory, not restricted. Let us call this idea *the Thesis of Unrestrictedness*. If we include this thesis in a theory of well-being, we get the idea that we cannot really determine what value a person's existence has for this person unless we take *all* her pleasures and sufferings into account.

(H3) is naturally explained by another essentially hedonistic idea, viz. (if we ignore the issue of duration) the idea that

(H4) All nonderivatively good experiences are good in virtue of their pleasantness only, and all nonderivatively bad experiences are bad in virtue of their unpleasantness only. That is, there are no other good- or bad-making features of experiences besides their pleasantness and unpleasantness: As far as the nonderivative value-for of our experiences is concerned, it is totally irrelevant how these experiences have originated, what their objects are, what beliefs they are based on, and the like.

If we combine this idea with (H2) and (H3), we get: An experience is nonderivatively good for a person if and only if (and because) it is pleasant, and an experience is nonderivatively bad for a person if and only if (and because) it is unpleasant.

This is how the pure hedonist answers **(II)** "To the extent that it is

possible: How do we determine just how (nonderivatively) valuable a certain pleasant (or unpleasant) experience is for a certain person?":

(H5) If we (again) ignore the issue of duration: The (positive or negative) value of an experience for the person who has it is a function of one thing only, viz. how pleasant or unpleasant it is. That is, the degree to which a pleasant experience is nonderivatively good for the experiencing subject is a function of its degree of pleasantness, and the degree to which an unpleasant experience is nonderivatively bad for the subject is a function of its degree of unpleasantness[52]. The function in question can be characterized as follows: The more pleasant an experience is, the (nonderivatively) better it is for an experiencing subject to have it: the higher is its prudential value. Or more specifically, it is always (nonderivatively) better for a person to feel pleasure than to suffer, it is always better for a person to have a more pleasant experience than to have a less pleasant experience, and it is always worse for a person to have a more unpleasant suffering than to have a less unpleasant suffering. This is the *"intensity-orientation"* of the hedonistic theory, and it follows naturally from the idea that the more there is of a good thing, the better. We can also assume that the degree to which a certain experience is good or bad for the experiencing subject is, on the hedonistic theory, *proportional* to how pleasant or unpleasant the experience is, i.e. that the marginal positive value-for of pleasure is not diminishing (and that the marginal negative value-for of suffering is neither diminishing nor increasing)[53].

---

[52]If we would not ignore the issue of duration, we would (instead) get the following idea: (H5') "The (positive or negative) value of an experience for the person who has it is a function of two things only, viz. (i) how pleasant or unpleasant it is, and (ii) how long it lasts (or how long it appears to last). That is, the degree to which a pleasant experience is nonderivatively good for the experiencing subject is a function of its degree of pleasantness and its duration, and the degree to which an unpleasant experience is nonderivatively bad for the subject is a function of its degree of unpleasantness and its duration". There will be a detailed discussion of this idea in appendix E, where I argue (in a somewhat "speculative" way) for the following two theses: (a) If duration matters, it is a certain kind of subjective duration that matters, and (b) it may well be the case that duration does not matter at all, and that we can (therefore) continue to ignore it.

[53]As I see it, it is (on the face of it) not implausible to assume that the positive value of a certain amount of pleasure is proportional to this amount, but is the assumption really well-founded? That is, does the "idea of proportionality" have any support; are there really any good reasons for assuming that the marginal

This is (roughly) how the pure hedonist thinks we should (to the extent that it is possible) determine how well off a certain person is (on the whole) at a certain time, i.e. this is the answer that he gives to **(III)**:

> **(H6)** The value that a certain life-at-a-certain-time has for the person who is living it is a function of how pleasant (unpleasant) its "experiential content" is at that time, i.e. of how much pleasure and suffering that it contains at the time. The more pleasure and the less suffering a life contains, the better this life is for the person who is living it. For example, if two possible lives-at-certain-times contain the same "amount" of suffering, but different "amounts" of pleasure, then it is better for a person to have the life that contains more pleasure.

It is important to point out that all hedonists are not pure hedonists, however: there are also hedonists who do not accept all the claims (H1)-(H6) above. Or alternatively put, there are also *"modified" versions* of the hedonistic theory. On my view, these modified theories are basically of two kinds: First, there is the type of hedonist who (like Mill) accepts (H1)-(H3) but rejects (H4)-(H6), and second, there is *the "restricted hedonist"*, who accepts (H1) and (H2) but rejects (H3)-(H6).

That is, all modified hedonists accept (H1), the Experience Requirement, and (H2), the idea that the only thing that can be nonderivatively good for a person is to have pleasant experiences (etc.). On my view, this is what makes the modified hedonist a hedonist. It is (H1) and (H2) that constitute "the essence of hedonism", and there is no way one can reject any of these claims and still consider oneself a hedonist.

Moreover, all modified hedonists reject (H4)-(H6). If we restrict our attention to the case of good experience and ignore the issue of duration: The modified hedonist denies that every nonderivatively good experience is good in virtue of its pleasantness only; he denies (for this reason) that the positive value of an experience for the person who has it is always proportional to how pleasant it is; he even denies that it is always the case that the more pleasant an experience is, the better it is for the experiencing subject to have it.

The modified hedonist replaces (H4)-(H5) with the following claim:

---

value of pleasure is not diminishing?

(MH4) If we ignore the issue of duration: There are other "value-for-affecting features" of experiences besides pleasantness and unpleasantness. The value of an experience for the person who has it is does not just depend on how pleasant or unpleasant it is, it also depends on what other (relevant) features the experience has; and if any of the relevant features comes in degrees, the value of the experience might also depend on to what degree it possesses the feature in question. For example, the value-for of a pleasant experience might (in part) depend on whether the belief on which it is based is true or false, or on whether it is associated with the employment of our "characteristically human capacities" or not.

But how exactly does the modified hedonist think the value-for of an experience depends on its "value-affecting features"? How should we (on this theory) calculate the value-for of an experience from information about its relevant descriptive features, i.e. how should the value function be characterized? Suppose that our modified hedonist claims that the nonderivative value-for of a pleasant experience is a function of two things only, viz. how pleasant it is, and whether it can be classified as "higher" or "lower" (or more specifically, that it is somehow nonderivatively better for a person to have pleasant experiences of the higher type). How exactly should this claim be interpreted? Well, this is (most probably) what our modified hedonist would say:

(a) First of all, it should be noted that (H2) tells us that an experience cannot be good unless it is pleasant. This puts certain restrictions on the function.

(b) If our modified hedonist is not a restricted hedonist (cf. below), he also accepts (H3), the idea that every pleasure is nonderivatively good for the experiencing subject, regardless of whether it is classified as higher or lower. This also puts certain restrictions on the function.

(c) If two pleasant experiences are equally pleasant, and if one of them is higher while the other is lower, then it is better for a person to have the higher pleasure (assuming that the two pleasures are comparable with respect to pleasantness).

(d) If two pleasant experiences are of the same type (e.g., if they are both of the higher type), then it is better for a person to have the experience that is more pleasant.

(e) Our modified hedonist would (most probably) *not* claim that every

pleasure of the higher type is better than every pleasure of the lower type, e.g., that a very pleasant pleasure of the lower type is less good than a much less pleasant pleasure of the higher type (again, assuming that the two pleasures are comparable with respect to pleasantness). Instead, he would make a weaker claim, viz. he would say that even though it is *ceteris paribus* better for a person to have higher pleasures than to have lower pleasures: If a lower pleasure is sufficiently pleasant it can outweigh a higher pleasure[54].

In short, our modified hedonist would make the following (imprecise) *claim concerning relative weights*: "If the pleasant experience E1 is of the higher type, and if the pleasant experience E2 is of the lower type, then it might be nonderivatively better for the experiencing subject to have E1, even if E2 is more pleasant than E1. And if E1 and E2 are equally pleasant, then it is better to have E1 than to have E2 (because E1 is higher while E2 is lower)". I don't think he can get any more precise than this.

So, would our modified hedonist make similar claims about sufferings and unpleasantness? That is, would he claim that it might be less bad to have a more unpleasant suffering because it is of a higher type? I think not. I haven't come across any explicit statements on this issue, but it is my guess that as far as suffering and unpleasantness is concerned, modified hedonists tend to be pure hedonists.

A modified hedonist might also reject (H3), however, i.e. he might be a Restricted Hedonist. The restricted hedonist replaces (H3) with the following claim:

(RH3) There are pleasant experiences that are not nonderivatively good for the experiencing subject (there might even be pleasures that

---

[54]As far as I can see, (MH4) can also be formulated in terms of organic wholes, viz. in the following way: A complex whole cannot have positive nonderivative value-for-P in isolation (i.e. positive *intrinsic value* in the proper sense of the term) unless it is of the type "P has a pleasant experience", and it cannot have negative intrinsic value unless it is of the type "P has an unpleasant experience". However, the intrinsic value of such a whole is not always a function of how pleasant or unpleasant its experiential content is. There are also other features, the presence or absence of which might affect the value of a whole. For example, it might be intrinsically better for a person to have an emotion that is both pleasant and based on a true belief than it is to have an emotion that is both pleasant (to the same degree) and based on a false belief, even though it not intrinsically valuable to have true beliefs. (If anyone is tempted to use the notion of contributory value here, cf. note 36).

are bad for the subject). It is also possible (but not likely) that there are sufferings which are not bad for the suffering subject. Or alternatively put, every restricted hedonist makes some kind of *restriction claim*, i.e. a claim of the following form: "If a pleasant experience has a certain "non-hedonic" feature F (e.g., if it is sadistic or based on a delusion), then it does not have nonderivative value for the subject to have it, in spite of its pleasantness"[55].

The fact that the modified hedonist accepts (MH4) (and maybe also (RH3)) means that he has to reject (H6). He would (instead) accept the following claim:

(MH5) The value that a certain life-at-a-certain-time has for the person who is living it is (roughly) a function of how much good pleasure and how much bad suffering that it contains (and so on).

The satisfaction interpretation of the Desire-Fulfilment Theory

There are a number of different versions of the desire theory. The only thing which these versions have (on the satisfaction interpretation) in common is that they all accept the following claim:

(D1) Nothing but (actual) desire-fulfilment can be nonderivatively good for a person: the only thing that is nonderivatively good for a person is the circumstance that his actual intrinsic desires are satisfied[56]. However, this does not allow us to assume that non-fulfilment (or frustration) of desire is nonderivatively bad for a person: from the fact that it is good for a certain person to have a certain desire satisfied, we can not conclude that it is bad for this person if the desire is not satisfied. As far as badness is concerned, we should

---

[55]A restriction claim for sufferings would take the form "If an unpleasant experience is of the kind K, then it is not nonderivatively bad for people to have it", but I doubt whether any such claims are plausible. I also doubt that there are pleasures that have negative value for the experiencing subject; it might have negative (nonderivative) *value-period* that a certain person feels sadistic joy, however.

[56]So, does this mean that so-called idealized versions of the desire theory are not really desire theories? Not really. All it means is that idealized theories are inconsistent with the satisfaction interpretation of the theory: all such theories presuppose the object interpretation, and they can (therefore) not be regarded as substantive good theories.

adopt the following view: Nothing but "aversion-fulfilment" can be nonderivatively bad for a person, i.e. the only thing that is nonderivatively bad for a person is (roughly) that some situation to which he has an aversion obtains.

This is all that the different versions of the (actual) desire theory agree on. This means that from now on, we have to treat different versions of the theory separately. Let start with the simplest version, viz. *the unrestricted (actual) desire theory*. Besides (D1), this theory makes the following claims:

(UD2) Every fulfilment of an intrinsic desire is nonderivatively good for the desiring subject, and every "fulfilment" of an intrinsic aversion is nonderivatively bad for the averse subject. That is, the class of intrinsic desires that it is nonderivatively good for a person to have fulfilled, and the class of intrinsic aversions that it is nonderivatively bad to have fulfilled, these classes are, on this version of the theory, not restricted. Let us call this idea *the Thesis of Unrestrictedness*.

It is (again) important to note that the theory makes no claims about what value it has for a person *not* to have his desires or aversions fulfilled. For example, the theory does not claim that it is bad for a person not to get what he wants, or that it is good for a person not to get what he does not want. Suppose that it is good for P to have his desire that X satisfied, and that it is bad for P to have his aversion to Y satisfied. This does not automatically imply that it is bad for P that not-X obtains, or that it is good for P that not-Y obtains. The only thing that determines what value not-X and not-Y have for P is (roughly) where on P's preference ordering these negatively specified facts can be found. Here, it should be observed that in real life, a person's desire that X is often accompanied by an aversion to not-X, and in such a case it is obviously bad for this person that X does not obtain. But again: the reason why not-X is bad for him is that he has an aversion to not-X, and not that he has a desire that X. What is bad for a person is "to get what he wants not to have", rather than not to get what he wants.

(D1) and (UD2) constitute the unrestricted desire theorist's answer to **(I)** "What kinds of local situations have nonderivative value for a person?". This is how he answers **(II)**, i.e. (in the present case) "How do we determine just how (nonderivatively) valuable a certain desire-

fulfilment or aversion-fulfilment is for a certain person?":

(UD3) The degree to which a certain desire-fulfilment is non-derivatively good for the desiring subject is a function of one thing only, viz. of how strong the desire is, and the degree to which an aversion-fulfilment is nonderivatively bad for the averse subject is (in a similar way) a function of how strong the aversion is. No other properties of our desires and aversions besides their strength are of any relevance in this context, e.g., it does not matter at all how they have originated, what their objects are, and so forth. The function in question can be characterized as follows: The "stronger" (the more "intense") an intrinsic desire is, the (nonderivatively) better it is for the desiring subject to have it fulfilled, and the stronger an intrinsic aversion is, the worse it is for the averse subject to have it "fulfilled". This is the *"intensity-orientation"* of the desire theory. We can also assume that the value it has for a desiring subject to have a certain desire or aversion fulfilled is (on the theory) *proportional* to how strong it is, i.e. that the marginal values of desire-fulfilment and aversion-fulfilment are not diminishing.

This is the answer that the unrestricted desire theorist gives to **(III)**, i.e. this is how he thinks we should (to the extent that it is possible) determine how well off a certain person is at a certain time:

(UD4) A person's level of well-being is dependent on what his intrinsic desires and aversions are, and whether these desires and aversions are fulfilled or not. Or more specifically, the more desire-fulfilment and the less aversion-fulfilment a life contains, the better this life is for the person who lives it, e.g., if two possible lives contain the same "amount" of aversion-fulfilment, but different "amounts" of desire-fulfilment, then it is better for a person to live the life that contains more desire-fulfilment.

This ends our characterization of the unrestricted desire theory. There are also a number of *restricted versions* of the desire theory, however. The central feature of these theories (besides the fact that they accept (D1)) is that they reject (UD2), i.e. the thesis of Unrestrictedness.

(RD2) Even though it is true that nothing but desire-fulfilment can be nonderivatively good for a person, and that nothing but aversion-fulfilment can be nonderivatively bad for a person (as (D1) claims): It

is not the case that every fulfilment of every intrinsic desire is nonderivatively good for the desiring subject, nor that every "fulfilment" of every intrinsic aversion is nonderivatively bad for the averse subject. Or alternatively, the class of intrinsic desires that it is nonderivatively good for a person to have fulfilled, and the class of intrinsic aversions that it is nonderivatively bad to have fulfilled, these classes are (according to these versions of the theory) both *restricted*.

Let us now introduce the distinction between *relevant and irrelevant (intrinsic) desires and aversions*. Let us say that an intrinsic desire is relevant if and only if it is nonderivatively good for the desiring subject to have it fulfilled (if its fulfilment makes him better off); that an intrinsic aversion is relevant if and only if it is nonderivatively bad for the averse subject to have it fulfilled (if its fulfilment makes his life worse); and that an intrinsic desire or aversion is irrelevant if and only if it is not relevant. If we adopt this terminology, we can see that (UD2) (the thesis of Unrestrictedness) states that *all* intrinsic desires and aversions are relevant, while (RD2), on the other hand, states that *only some kinds* of intrinsic desires and aversions should count as relevant. Or alternatively, every restricted desire theory makes restriction claims of the following form: "Only desires and aversions of a certain type should count as relevant. If an intrinsic desire (or aversion) is not of this type, it can never be nonderivatively good (or bad) for the subject to have it fulfilled".

There are many different versions of the restricted desire theory, and what makes these versions different is (of course) that they have different ideas about exactly how the class of (all) intrinsic desires (and aversions) should be restricted. Some restricted theories are *"object-oriented"*: they make restriction claims of the form "only desires and aversions with certain kinds of objects (e.g., only desires that are about one's own life) should count as relevant; if a desire (aversion) has the wrong kind of object, it can never be good (bad) for the subject to have it fulfilled". Other restricted theories are *"rationality-oriented"*; these theories make restriction claims of a different form, viz. "only rational (e.g., informed) desires and aversions should count as relevant; it can never be good (or bad) for a person to have an irrational desire (or aversion) fulfilled". There are also restricted theories that are neither "object-oriented" nor "rationality-oriented".

This was the restricted desire theorist's answer to the non-comparative **(I)**. So, how does he answer the comparative questions **(II)** and **(III)**? On the assumption that we restrict our attention to the relevant desires and aversions: How do we determine just how valuable a certain desire-fulfilment or aversion-fulfilment is for a certain person, and how do we determine just how well of a certain person is at a certain time?

As I see it, there is nothing that prevents a restricted desire theorist from being "intensity-oriented", i.e. he may well accept (UD3) above. Or more precisely, he can (so to speak) be intensity-oriented after the restriction has been made, e.g., he can think of a restriction as nothing but an elimination of certain items from our intrinsic preference orderings, and he can claim that this is all we should do: once we have eliminated all irrelevant items from these orderings, we should let the remaining (relevant) items maintain their relative positions in the orderings. That is, as far as the relative importance of the *relevant* desires and aversions is concerned, strength is all that should count.

> **(UD3')** How good (or how bad) it is for a desiring subject to have a certain *relevant* desire (or aversion) fulfilled depends on one thing only, viz. how strong the desire (or aversion) is. That is, once we have eliminated certain desires and aversions as irrelevant, strength is the only property of our desires and aversions that counts. For example, it is always (nonderivatively) better for a desiring subject to have the stronger of two relevant intrinsic desires fulfilled.

An intensity-oriented restricted desire theorist would also replace (UD4) with

> **(UD4')** A person's overall level of well-being is a function of how much good desire-fulfilment and how much bad aversion-fulfilment there is in his life, i.e. it is a function of how many *relevant* desires and aversions that are fulfilled, and of how strong these desires and aversions are.

However, a desire theorist need not accept (UD3') or (UD4'): it is also possible to reject the intensity-orientation (and notice that even though such a move is "restrictive" in spirit, it does not presuppose that (RD2) is accepted). This is what such a "non-intensity-oriented" desire theory would claim:

(NID3) How good (or bad) it is for a desiring subject to have a certain relevant desire (or aversion) fulfilled does not just depend on how strong it is, but also on what other properties it has, e.g., on whether it is rational or irrational, on what content it has, or the like. Strength is not all that counts: it is not always nonderivatively better for a desiring subject to have the stronger of two relevant intrinsic desires fulfilled, and the reason for this is that it is *ceteris paribus* better for a person to have desires of certain types (e.g., autonomous desires) fulfilled than to have desires of other types (e.g., heteronomous desires) fulfilled.

If a desire theorist accepts (NID3) rather than (UD3'), he also has to reject (UD4'), i.e. the idea that the value of a life for the person who lives it is a function of how much desire-fulfilment and aversion-fulfilment this life contains. But it seems (to me) impossible to find an intelligible formulation that can replace (UD4').

This ends our "positive" characterization of (the satisfaction interpretation of) the desire theory. We will soon give a negative characterization of the same theory, viz. by contrasting it with hedonism. But before we do this, let us first, for the sake of completeness, mention that there are also so-called *idealized versions of the desire theory*. These theories are (on my view) not consistent with the satisfaction interpretation of the desire theory, and they can not be regarded as substantive theories (cf. note 56). Instead, they are formal theories that presuppose the object interpretation of the theory. To get a rudimentary grasp on what the idealized desire theory is about, let us contrast it against the object interpretation of the unrestricted theory.

If the we replace the satisfaction interpretation by the object interpretation, the conjunction of (D1) and (UD2) has to be reformulated as follows:

(OD1) A situation X is nonderivatively good for a person P if and only if (and because) P has an (actual) intrinsic desire that X holds, and a situation Y is nonderivatively bad for P if and only if (and because) P has an (actual) intrinsic aversion to Y[57].

An idealized desire theory rejects this idea, but also

---

[57]This formal criterion is a conjunction of (D1) and (D2) in appendix D.

(OD2) The nonderivative value of a situation X for a person is a function of one thing only, viz. how desirable or undesirable X is for this person. For example, if X and Y are both desired by a certain person, it is nonderivatively better for this person that X obtains than that Y obtains if and only if (and because) his desire that X is stronger than his desire that Y.

Instead, it makes the following claims:

(ID1) A situation X is nonderivatively good for a person P if and only if (and because) P *would* intrinsically desire that X under ideal circumstances (e.g., if he were fully rational, had all the relevant information, were thinking clearly, or the like), and a situation Y is nonderivatively bad for P if and only if (and because) P *would*, under ideal circumstances, have an intrinsic aversion to Y.

(ID2) The nonderivative value of a situation X for a person P is a function of one thing only, viz. how desirable or undesirable X *would* be for P under ideal circumstances. For example, it is nonderivatively better for a person that X obtains than that Y obtains if and only if (and because) he *would* (under ideal circumstances) intrinsically prefer X to Y.

This ends our positive characterization of the desire theory. Let us now take a quick look at how it differs from the hedonistic theory.

The desire theory vs. hedonism

From all that has been said above, we might (it seems) conclude that as far as nonderivative value-for is concerned, there is a rather strong disagreement between the desire theory and the hedonistic theory. For example, one might believe that there is no "overlap" at all between the class of situations that are (on the desire theory) nonderivatively good and bad for a person, and the class of situations that are (according to the hedonistic theory) nonderivatively good and bad for a person. As a rule, this is true, but it is important to notice that there are two possible exceptions to this rule (two possible ways in which the two theories might, in part, coincide):

First, most of us have an intrinsic desire to feel good and an intrinsic aversion to feeling bad, and to the extent that we have such preferen-

ces, the object interpretation of the desire theory implies that it is nonderivatively good for us to feel pleasure, and nonderivatively bad for us to suffer. And second, we have not yet said anything about the nature of pleasantness and unpleasantness, but one of the dominant views on this matter is the so-called "preference-hedonist" view. On this view, an experience is (roughly speaking) pleasant for a person if and only if (and because) it is (in a certain way) desired by this person, and an experience is unpleasant if and only if (and because) the experiencing subject has a certain kind of aversion to it. If this is correct, it seems that all pleasant and unpleasant experiences can be viewed as fulfilments of a certain kind of "now-for-now" desires and aversions, and that the hedonistic theory can (for this reason) be regarded as a special case of the desire theory, viz. as a strongly restricted (object-oriented) version of this theory. But again, this is only true on the object interpretation of the desire theory.

This suggests that there is no "extensional overlap" at all between the satisfaction interpretation of the desire theory and the hedonistic theory. If fact, there is (normally) not much of an overlap between the object interpretation of the desire theory and the hedonistic theory either, e.g., we can safely assume that many of the situations that are (on the object interpretation of the desire theory) nonderivatively good for a person are *not* regarded as nonderivatively good by the hedonist, and (perhaps) vice versa[58]. This is (of course) mainly due to the fact that hedonism is an "internalist theory" (it only attributes nonderivative value-for-P to intrinsic facts about P), while most versions of the desire theory are (on the object interpretation of this theory) "relationalist theories"[59] (they also attribute nonderivative value-for-P to relational facts about P); there are even desire theories which can be regarded as "externalist theories" (they even attribute nonderivative value-for-P to facts that are external in relation to P). As Scanlon (1993) points out,

[h]edonism takes certain mental states to be the only things of ulti-

---

[58]Or more specifically; all situations that are, according to the *preference-hedonist*, nonderivatively good for a person are also regarded as nonderivatively good by the object interpretation of the desire theory; however, there may well be situations that are regarded as nonderivatively good by the *quality-hedonist* but not by object interpretation of the desire theory. Cf. section 2.2.

[59]The relationalist theory *par excellence* is of course the satisfaction interpretation of the desire theory.

> mate value. Desire theories [here, Scanlon takes (like almost everybody else) the object interpretation for granted] count things as valuable if they are the objects of certain "mental states" or attitudes [i.e. desires], but the things valued need not be mental states and the attitudes which confer value need not themselves be valuable (p 189).

Or alternatively, "Desire Theories [are likely to] reject the experience requirement and allow that a person's life can be made better and worse not only by changes in that person's states of consciousness but also by occurrences elsewhere in the world which fulfil that person's preferences" (ibid., p 186).

However, these differences between the hedonistic theory and the object interpretation of the desire theory are consistent with the fact that they often agree on what particular situations that are good and bad (derivatively *or* nonderivatively) for a person[60]. The reason why they tend to agree on what particulars that are good and bad for a person is simple: To have one's desires fulfilled is often pleasant, and to have one's aversions fulfilled is often unpleasant. One might even believe (e.g., with Epicurus) that most of our pleasant experiences are based on (or caused by) desire-fulfilment, and that most of our suffering is based on (or caused by) "aversion-fulfilment". However, this is only true on condition that the person's beliefs about the world are (in the relevant respects) true, i.e. it would be more correct to say that many pleasant experiences are based on (or caused by) some *belief* (true or false) that some desired situation obtains (and similarly for sufferings). This means that the two theories can not be expected to agree about what particular facts that are good and bad for a person when the person's beliefs about the world are (in the relevant respects) false.

Now that we are familiar with the possible sources of disagreement (about particular cases) between the hedonist and the desire theorist, let us (now) take a look at what these disagreements might look like. There are at least four types of cases here:

(1) First we have the "weakest" type of case, where the two theories agree on that a particular desire-fulfilment is good for a person, but they disagree on how good it is. These are the cases where the satisfaction of a desire is accompanied by a pleasant experience, but where

---

[60]From now on, and until the end of this section, it is no longer important which of the two interpretations of the desire theory we have in mind.

the magnitude of the pleasantness does not "correspond to" the strength of the desire fulfilled.

(2) There may also be situations which (on the desire theory), make a life better or worse, but where the Hedonistic Theory is "indifferent". These are the cases where the satisfactions of our desires or aversions are not accompanied by any pleasant or unpleasant experiences at all. The most interesting cases of this type are the cases where some preference is, unknown to the desiring subject, satisfied. Suppose that my desire that my American friends are happy, or my aversion to being deceived, is satisfied without my knowing it. In this case, most versions of the desire theory would claim that this makes my life better (or worse) than it would otherwise have been, while the hedonist would regard the circumstance as irrelevant. This follows from the idea that "what you don't know can't hurt you (or benefit you)", an idea that (somehow) captures the essence of hedonism.

(3) The disagreement between the two theories may be even stronger than this, viz. if there are cases where something is, on the desire theory, good for as person, while the hedonistic theory claims that it is bad for this person. For example, it may (on the desire theory) be in a person's interest to have a strong intrinsic desire fulfilled, even if this makes him less happy (cf. Parfit (1984), p 465). Kekes (1988), who is himself a desire theorist, gives the following examples:

> /.../ the satisfaction of wants, essential to good lives, need not involve episodes of feeling happy, necessary for happiness. The wants whose satisfaction we rationally seek may concern doing hard, unpleasant, altruistic, impersonal, dutiful, self-sacrificial, heroic, self-denying, and occasionally even self-destructive things. And none of these is usually conducive to feeling episodically happy (p 167).

(4) There may also (if the preference-hedonism is wrong about the nature of pleasure and displeasure) exist situations which the hedonist regards as making a person's life better or worse, but which the desire theorist conceives of as irrelevant. One possible example of this are the cases where someone is happy because he believes that some desired situation obtains, but where his belief is false. Another example is physical pain, which (it might be argued) is simply unpleasant, regardless of whether or not the person who is in pain has an aversion to being in pain.

This concludes the introductory sections on the first two types of theories of prudential value, i.e. hedonism and the desire theory. Let us now move on to the third type of theory listed by Parfit (1984), viz. the objective list theory (viewed as a substantive evaluative theory).

"The Objective List Theory", or "non-internalist pluralism"

This is how Parfit (1984) characterizes the Objective List Theory:

> According to this theory, certain things are good or bad for people, whether or not these people would want to have the good things, or to avoid the bad things. The good things might include moral goodness, rational activity, the development of one's abilities, having children and being a good parent, knowledge, and the awareness of true beauty. The bad things might include being betrayed, manipulated, slandered, deceived, being deprived of liberty or dignity, and enjoying either sadistic pleasure, or aesthetic pleasure in what is in fact ugly (p 499).

Now, if we want to derive a plausible characterization of the objective list theory *qua* substantive evaluative theory from this passage, we have to focus on the second part, i.e. on all the things that such a theory *might* conceive of as nonderivatively good and bad for a person (the first part of the characterization is not really a description of the theory *qua* substantive evaluative theory; cf. appendix D). Now, if we do this, it is (I think) rather clear that the theories that Parfit has in mind are those universalist[61] (or general) substantive good theories that are both **(1)** pluralist theories, and **(2)** what Kagan (1992) calls externalist theories, and what may also be called relationalist theories[62]. So, let us (therefore) take a look at what the terms "pluralism", "externalism", and "relationalism" mean in this context.

Monism vs. pluralism

The hedonistic theory and the desire-fulfilment theory (on the satisfaction interpretation) are both *monist theories*: they both claim (or imply) that all (local) situations that are nonderivatively good for a person are

---

[61]This notion is explained elsewhere, e.g., in appendix D.
[62]Cf. note 51 above.

of one and the same type (i.e. pleasures or desire-fulfilments), and that all situations that are nonderivatively bad for a person are of another type (i.e. sufferings or aversion-fulfilments). The objective list theories, on the other hand, are *pluralist theories* of prudential value: they claim that the situations that are nonderivatively good for a person are of several different types, and that this holds for the nonderivatively bad situations as well.

So, why is it so obvious that objective list theories are pluralist theories? Well, one reason is that if these theories were not pluralist, they could not be (plausibly) contrasted with hedonism. It seems that every plausible version of the objective list theory must accept the idea that it is (at least sometimes) nonderivatively good for a person to feel pleasure, and bad for a person to suffer, and this implies that a plausible objective list theory can not really be contrasted with hedonism unless it is a pluralist theory.

That we should regard objective list theories as pluralist theories is also indicated by the term "list", and by the fact that these theories claim (according to Parfit) that certain things, in the plural, are good or bad for us. As the term "list" suggests, every objective list theory provides us with two lists of "things" (facts, situations), one "positive list" and one "negative list". The items to which the positive list refers are the different "things" which are, according to the theory, nonderivatively good for us, i.e. facts which, if they obtain, make us better off than we would otherwise have been (and similarly for the negative list). The types of facts referred to on the two lists may (of course) be internal to the person, but they can also be relational facts about the person, or external in relation to the person, or even external in relation to the person's life. (If we look at Parfit's examples, however, it is obvious that the emphasis is on relational facts).

It should be noted that the object interpretation of the desire theory is also a pluralist theory, viz. a subjectivist (and relativist) pluralist theory. That is, a pluralist theory of prudential value can be both subjectivist and objectivist, and both "universalist" and "relativist". Universalist versions of pluralism assume that our respective lists are (given a general enough description) identical, while relativist versions of pluralism claim (instead) that our respective lists differ, i.e. that "what is good for me (or us) may not be good for you (or them)". It is likely that such relativist ideas are based on some kind of subjectivism, but it

is also possible that a theory is both objectivist and relativist (cf. appendix D). The objectivist theories with which we will be concerned in this book are all universalist, however.

Let us now look at the second central feature of the objective list theories, i.e. the fact that they are "externalist" or "relationalist" theories, or in short, "non-internalist" theories.

Internalism, relationalism, and externalism

As far as I can tell, the distinction between *internalist and externalist theories* of prudential value was introduced by Kagan (1992) (who thinks that the distinction in question is the most interesting distinction between different substantive theories of prudential value (well-being). This is how he characterizes the difference:

> /.../ [T]he more fundamental distinction seems to be between theories that limit well-being to intrinsic facts about the person and theories that allow for relational facts to directly contribute to well-being as well. Mental state theories [i.e. theories "which hold that an individual's well-being consists solely in the presence of the relevant kinds of mental states" (cf. p 169)] may be the most well-known or the most plausible examples of the former type of theory, but they do not exhaust the class. Perhaps this /.../ division should be labeled as the division between intrinsic theories and relational theories (or intrinsic theories and extrinsic theories), but I myself am drawn to a slightly different set of labels: some theories restrict well-being to facts internal to the person; other theories allow for the direct relevance of facts external to the person as well (p 188).

That is, a theory of well-being is internalist (in Kagan's terminology) if and only if it claims that all facts that are nonderivatively good and bad for a certain person are "internal to this person", or alternatively, that a person's level of well-being depends solely upon "facts internal to this person" (physical or mental). A theory of well-being is externalist, on the other hand, if it rejects this idea, i.e. if and only if it makes a person's level of well-being directly dependent on not just on intrinsic facts about the person, but on other kinds of facts as well, viz. on "facts external to the person", or on "relational facts".

On my view, this is a problematic distinction, and the reason for this is

that it is not quite clear what facts that Kagan has in mind when he talks about facts about the person. So, how should we distinguish facts about persons (or descriptions of persons) from other kinds of facts (or descriptions)?[63] For example, can (and should) a fact that is external to a person ever count as a fact about this person?

It can hardly be doubted that "P has a big nose" and "P is generous" are descriptions of P. After all, these are descriptions by means of which some "more permanent" bodily or mental feature is attributed to P, descriptions which tell us what *kind* of person P is. It also seems clear that descriptions like "P is depressed", "P has a fever", and "P feels pleasure" are descriptions of P. These descriptions do not tell us what kind of person P is, however, and we might (therefore) think of them as state-descriptions of P-at-*t* rather than descriptions of P (as a person).

The fact that P's desire to have children is fulfilled, the fact that P is married to Q, and the fact that P owns a sailing boat are all relational facts which involve P, and even though these facts are not (strictly speaking) *about P*, there is no reason why we should not (in this context) think of them as such.

In a similar way, we can think of action-facts, activity-facts, and behavioural facts as facts about persons (e.g., *qua* agents). However, descriptions like "P is writing a book" and "P is scratching his (big) nose" are more easily regarded as descriptions of *P's life* than as descriptions of P[64]. This makes it rather difficult to determine whether the facts to

---

[63]There is also another reason why we might want to know what kinds of facts which are included in the category "facts about a person", viz. if we accept Kagan's claim that one of the conditions that an "adequate theory of well-being [i.e. a "theory [that] /.../ attempts to specify in general terms the set of facts that comprise the good for the individual"] would have to meet" is the condition "that the specified facts must be *about the person*" (ibid., p 185, my italics).

[64]Descriptions of P are typically of the form "P has the feature F", where F is (like P himself) atemporal, i.e. it persists through time, but it is not (like an event) extended in time. But when we say about P that he is writing a book, or that he is scratching his nose, or that he feels pleasure, then we are not attributing any atemporal feature to P. Instead, we make a reference to something temporally extended (e.g., a mental event or an activity), and this suggests that we are describing P's life rather than P. But the reason why it is more appropriate to regard the fact that P is writing a book as a fact about P's life rather than as a fact about P is not just that actions are (by definition) temporally extended. The important thing in this context is that actions (etc.) are temporal wholes that can not be reduced to a mere sequence of states; this is why the fact that P is involved in the performance of a certain action at *t* is *not* (strictly speaking) a fact about P-

which these descriptions refer should be regarded as intrinsic or relational, but this is one way in which it might be done: We should conceive of an action-fact as relational if and only if it (its description) includes an implicit reference to something external, e.g., to an external result (as in the case of murder) or to certain conventions or social systems (as in the case of voting or writing a cheque). This suggests that all purely "behavioural facts" (like the fact that someone scratches his nose) and at least some action-facts (like the fact that someone is walking) can be regarded as "intrinsic".

The fact that P's wife has an affair with another man, and the fact that P's son is harassed in school, are both external in relation to P, and they can (for this reason) hardly be regarded as facts about P. They might, however, be regarded as facts about P's life, viz. on the assumption that P's wife and P's son are (in some sense) parts of P's life. The fact that the mountain gorilla survives as a species, or the fact that P's post-mortem reputation is excellent, on the other hand, are not just external in relation to P, but also in relation to P's life.

Now, this shows (I think) that Kagan's distinction between internalist and externalist theories is rather vague, and that it needs to be modified. So, I will now propose a slightly different classification of substantive theories of prudential value. This classification is (like every plausible classification) based on the idea that a substantive theory of prudential value "attempts to specify in general terms the set of facts that comprise the good for the individual" (cf. ibid., p 185), or alternatively, that it attempts to specify the set of facts of which it is true that it is nonderivatively good and bad for a person that they obtain. Different substantive theories of prudential value specify this set in different ways, and they can (therefore) be classified on basis of how they do this. The classification I want to propose is based on the distinctions between the different kinds of facts above, and the views which I will consider are (i) internalism, (ii) relationalism, (iii) person-externalism, and (iv) life-externalism.

(i) Internalism

In Kagan's terminology, internalist theories claim that the only facts that can be nonderivatively good and bad for a person are "facts internal to

---

at-*t*.

the person" or "intrinsic facts about the person"[65]. Or if we borrow some terminology from Nagel (1970), we can say that an internalist is a person who claims that all the goods and evils that can befall a man are restricted to nonrelational properties ascribable to him at particular times (cf. p 6). That is, an internalist can claim that it has nonderivative value for a person to be in a certain state (bodily or mental), to have a certain kind of body, or to be a certain kind of person, but he can (as I see it) also allow for the fact that certain kinds of "internal" events can be nonderivatively good and bad for a person, e.g., different kinds of mental events, processes, or activities.

Now, it is not very common that nonderivative value (positive or negative) is attributed to the fact that a person is in a certain bodily state, or that he has a certain kind of body, or that he is a certain kind of person, or that he is behaving in certain ways. In short, the most common form of internalism is "mental statism": the type of theory according to which the only facts that can have nonderivative value for a person are certain kinds of mental facts, e.g., that he is in a certain mental state, or that a certain kind of mental process takes place in him.

(ii) Relationalism

On the relational view, "there are goods and evils which are irreducibly relational", i.e. there are relational facts about us that have nonderivative value for us, and how well off a person is at *t* depends (in part) upon how he is (at *t*) related to things (e.g., other people) external to himself. Some of the relational facts which are (supposedly) good for P are "state-facts" of the form "P stands in relation R to X (where X is external to P)", but there are also a number of action-facts, activity-facts, and interaction-facts that are best regarded as relational. That is, a relationalist can claim that it has nonderivative value for a person to be married or to belong to a certain group, but he can also claim that it is good for a person to perform actions of certain types, or to be engaged in certain kinds of activities[66].

---

[65]This internalism has (of course) nothing to do with the metaethical (value theoretical) internalism that is discussed in appendix B.

[66]It seems plausible to assume that most of the relational facts to which the relationalist attributes nonderivative value are "internal to" (or part of) a person's life. (Assuming that Nagel (1970) is correct when he claims that "a man's life includes much that does not take place within the boundaries of his body and his mind"). This suggests that most existing versions of relationalism (i.e. "person-

(iii) Person-Externalism

The "person-externalist" attributes nonderivative value-for-P to facts that are external in relation to P (but not necessarily in relation to his life), e.g., he might claim that the value of a person's life at *t* is (in part) dependent on what properties certain external things have at *t*, or how certain external things are related to each other at *t*, or what external events are going on at *t*. Suppose that the male person P has a wife (Philippa) and a child (Bernard) whom he both loves. Suppose also that Philippa has an affair with another man, and that Bernard is harassed in school. Now, the idea that these facts are nonderivatively bad for P, even if he never finds out about them, is (I think) a good example of a person-externalist view (which is not necessarily life-externalist)[67].

Just as the hedonist theory is a good example of an internalist theory and the satisfaction interpretation of the desire theory is a good example of a relationalist theory, so the object interpretation of the unrestricted desire theory is an example of a theory which may well make (in particular cases) substantive claims of a person-externalist kind. In fact, it may even make substantive claims of a life-externalist kind.

(iv) Life-externalism

The "life-externalist" claims that facts which are external in relation to a person's life may have nonderivative value for this person. On this view, a person's well-being may be affected by states or events that (to speak with Nagel) "does not take place within the boundaries of his life". Now, the boundaries of a person's life are of two kinds; they are either temporal or "atemporal" ("spatial" is not a good term here), and depending on what type of boundary we have in mind, we get different versions of "life-externalism".

The atemporal version of "life-externalism" claims that the value of a person's life at *t* may (in part) depend on facts that are external to his

---

relationalism") are (at the same time) versions of "life-internalism".

[67]I assume that the two facts are really external to P, i.e. that they are not relational facts about P. Now, as I see it, this is compatible with the idea that it can only be bad for P that a certain fact obtains if P is (in some relevant way) related to this fact. The idea that it can only be bad for P that Philippa "deceives him" if he is in a certain way related to this fact (e.g., if he has an aversion to being deceived), and/or to Philippa, does not imply that the fact that she has sex with other men is a relational fact about P (it may count as a fact about his life, however).

life (facts that obtain at t), e.g., on what properties certain external things have at t, or how certain external things are related to each other at t. As has already been suggested, the object interpretation of the unrestricted desire theory seems to be of this kind. If I have a strong intrinsic desire that the mountain gorilla survives (as a species), then it is (on this theory) nonderivatively good for me that this happens, even if it I do not know about it, and even if mountain gorillas are not part of my life.

The temporal version of "life-externalism" claims that the value of a person's life (but not its value at any particular time!) may (in part) depend on what happened before he was born or what will happen after his death (events that are outside the temporal boundaries of his life). It should be noted that this is an existing view, e.g., there are philosophers (like Aristotle and Nagel) who have claimed that a person's well-being can be affected by post-mortem events[68].

The Objective List Theory revisited

Now that the distinctions between (i) monistic and pluralistic theories, and (ii) internalist, relationalist, person-externalist, and life-externalist theories, have been made, it is easy to see that all theories which has "traditionally" been classified as objective list theories (the theories which have been regarded as alternatives to both hedonism and the desire theory) are **(1)** pluralist theories, and **(2)** relationalist (and perhaps also externalist) theories. *Qua* pluralist theories, they differ from hedonism and the satisfaction interpretation of the desire theory (which are both monistic theories), and *qua* relationalist theories, they differ from hedonism (which is an internalist theory), but not from the satisfaction interpretation of the desire theory (which is a relationalist theory). This means that relationalist objective list theories can only be contrasted against the desire theory if we take them to claim **(3)** that there are other relational facts (besides desire-fulfilment and aversion-

[68]This is of course inconsistent with my idea that all goodness-for-P is goodness-for-P-at-t (cf. appendix C). The assumption that the welfare subject is a temporally located person-at-a-certain-time simply makes Nagel's (1970) idea that a person's well-being might be affected by how he is related to "circumstances which may not coincide with him either in space or in time" too absurd. (How could the claim that the value of P's life at t_1 is (in part) dependent upon how he is related to something external that happens at t_2 have any plausibility?).

fulfilment) that are nonderivatively good and bad for us. But as far as I can see, it need not reject the idea that desire-fulfilment is one of several relational facts which have value for a person[69].

This - i.e. **(1)**, **(2)**, and **(3)** - is the objective list theorist's answer to **(I)** "What kinds of local situations have nonderivative value for a person?". This is a very unspecific "answer", however, and the reason for this is that "non-internalist pluralism" is not really a substantive evaluative theory, but a *type* of substantive theory. In fact, the different versions of the "theory" need not ("substantively speaking") have anything in common. Or alternatively put, although every concrete version of the "theory" makes specific claims about what is nonderivatively good and bad for us, no version of the "theory" makes any specific substantive claims *qua* non-internalist pluralist theory. This means that no "objective list theory" can really be assessed *as such*. Strictly speaking, there is no such thing as an objective list *theory*, of which there are different versions, and it is (therefore) more appropriate to refer to these "versions" as *theories* (in the plural). It is these concrete theories (or better: the specific substantive evaluative claims which constitute these theories), and nothing else, that can be assessed[70].

As far as **(II)** and **(III)** are concerned, it is worth noting that the non-internalist pluralist theory does (as far as I know) not make any clear explicit claims about how we should determine how well off a certain person is (on the whole), or about how we should determine just how valuable a certain local situation is (e.g., how we should compare, with respect to value-for, a certain "amount" of one good with a certain "amount" of another good). However, we can make certain assumptions about how our pluralist would answer these questions. To begin with, we can (I think) safely assume that he would give the following (vague) answer to **(III)**:

(4) The prudential value of a person's life at t is somehow dependent

---

[69]It is worth pointing out that as far as I know, there is no pluralist who has included desire-fulfilment in his or her list of goods. The reason for this is probably that almost everyone has taken the object interpretation of the desire theory for granted.

[70]In this respect, it differs from hedonism and the satisfaction interpretation of the actual desire theory. There are (as we have seen) several versions of two latter theories as well, but all these versions share a common substantive content, and it is therefore much more appropriate to regard these two theories as real substantive theories.

on "how much" of the different prudential values that this life "contains" at t, or alternatively, of how well he is doing in a number of specified dimensions (areas, or domains). Or more specifically, the more a person "possesses" of these values, the higher is his level of well-being.

Now, this suggests that we cannot determine how well off a person is (on the whole) unless the following two conditions are met: (i) We can determine how much he "possesses" of the different values (how well off he is in the different dimensions), and (ii) we can compare (with respect to value-for) a certain amount of one "value" with a certain amount of another "value", i.e. we can (so to speak) make inter-dimensional comparisons with respect to prudential value. But how are we (on the assumption that it is possible) to do this? To the extent that it is possible, how do we determine (in principle) how well off a certain person is in a certain dimension, and how are we to compare (with respect to value-for) a certain amount of one "value" with a certain amount of another "value"? (This is what **(II)** looks like in this pluralist context). Well,

(5) On the pure version of the "theory", we must not appeal to people's preferences[71].

(6) Concerning the issue on how we should compare (with respect to value-for) a certain amount of one "good" with a certain amount of another "good": Here, we can (I think) safely assume that there is no value which *trumps* (or is lexically prior to) any other value (cf. Griffin (1986), p 83), i.e. that there are no two prudential values such that any amount of the one value, no matter how small, is more valuable than any amount of the other value, no matter how large.

(7) Concerning the issue on how we should compare (with respect to

[71]To allow that we appeal to preferences in this context is to accept a kind of mixed theory of well-being, e.g., to be an objectivist about the good and the bad (about what the relevant dimensions are) and a subjectivist about how these dimensions should be weighed against each other. This type of mixed theory seems to be rather common in the context of "quality of life measurements", where it is often assumed that a person's overall quality of life is a function of how well he is doing in a number of "pre-determined domains", but where the subject himself is (so to speak) allowed to determine the relative importance of the different domains, e.g., to rank the domains with respect to their "contributive value".

value-for) different amounts of the same "good": We must *not* assume that the prudential value of a certain valuable thing is proportional to how much there is of the thing. That is, we must allow for the possibility that there are prudentially valuable things such that the marginal value of the thing is diminishing[72]. For example, in the case of friendship, we must not assume that how valuable different amounts of friendship are for a certain person is proportional to how large these amounts are.

Should we also allow for the possibility that there are prudential values the marginal value of which is (in certain ranges) zero, or even negative? That is, should we allow for the possibility that there are prudential values which are not "the-more-the-better"-values? For example, is it plausible (for someone who regards friendship as a prudential value, that is) to reject the idea that the more friendship there is in a person's life, the better-for-him?

To make this idea more precise, let us first distinguish between two senses of "more", the quantitative sense and the qualitative sense. I don't think our pluralist would deny that if the quantity of friendship (e.g., the amount of time that one spends in the company of friends) is held constant, then the more friendship (in the qualitative sense), the better. So the question is: Is there any reason to deny that "the more friendship (in the quantitative sense), the better"? Well, it may seem so[73], but I think not. More generally, I think we can assume that all the

---

[72]That a certain good (e.g., friendship) has *diminishing marginal value* simply means that the more there is of the good, the less is the increase in value that a certain addition of the good gives rise to. It is worth noting that this kind of talk does not really make sense unless both the good (e.g., friendship) and the value of the good are measurable on interval scales, which suggests that the talk in question is probably "senseless".

[73]E.g., for the following two reasons: (i) "There are (as we know) other valuable things in life besides friendship. Now, it is highly likely that if a person who would spend all his time in the company of friends would suffer in other domains. So it is reasonable to believe that after a certain point has been reached, the more friendship there is in a life, the less there will be of other valuable things. Therefore, it is not the case that the more friendship, the better". (ii) "It seems plausible to assume that final value should primarily be attributed to a person's social life (or life as a whole) viewed as an "organic whole". Moreover, the value of this whole is (supposedly) not simply a function of the values of the ingredients (like friendship, love, more superficial relationships, and solitude), but also has to do with how these elements are balanced with each other. This suggests that the value of friendship is primarily "contributive", and this would (in turn) imply that it can not be assumed that "the more friendship, the better" (if something has

things that our pluralist regards as prudential values are "the-more-the-better"-values.

This is (roughly) what a pure non-internalist pluralist would (or should) claim.

1.3. On "method": Some epistemological assumptions

There are (roughly speaking) two sets of questions concerning the justification of value-for-claims, viz. (i) epistemological questions of a more general nature, e.g., "Assuming that there are valid value-for-statements; is it possible to justify such statements?" and "If so, in what general way (or ways) can this be done?", and (ii) the more "substantive" (or specific) "Assuming that valid value-for-statements can be justified, how exactly can (and should) this be done?"[74]. In this essay, the emphasis will be on the more "substantive" questions, i.e. there will be little or no discussion of the epistemological questions. I am going to take certain answers to these questions for granted, however.

My epistemological assumptions: Weak cognitivism

First, I will assume that it is possible to justify valid value-for-statements (and value-for-beliefs)[75]. That is, I will assume that what Bergström (1990) calls "cognitivism" is true in this area, where cognitivism is under-

---

contributive value only, then we can not assume that the more there is of this thing, the better)".

These are both bad arguments, however. The first ("pragmatic") argument does not show that less friendship can be (*other things being equal*) better than more friendship. And the second argument claims that what has prudential value is not friendship as such, but certain types of social wholes (where it may well be the case that such a whole can not have value unless it contains a certain amount of friendship, however).

[74]Where subjectivism and objectivism (cf. appendix D) should be viewed as answer to the second question, and *not* to the first.

[75]Here, we should remind ourselves that the claims which are of primary importance in this context are the claims made by general (substantive) theories of prudential value, viz. claims of the form "all facts of type X have nonderivative value for all human beings (at all times)". However, the idea is not just that these general claims can be justified, but also that it is possible to justify both particular claims of the form "the particular fact X is nonderivatively good (or bad) for a particular person P" and "semi-general" claims of the form "all facts of type X have nonderivative value for a particular person P".

stood narrowly (as a conception of justification) rather than broadly (as a conception of knowledge in the traditional sense, where knowledge implies truth, in some "non-minimalist" sense of the term)[76]. The assumption is that rational argument is possible in this area, i.e. that there is such a thing as having good reasons for accepting or rejecting a certain theory of prudential value. I will not assume that *all* valid value-for-statements can be justified, however: the cognitivism I take for granted is "weak" rather than "strong". There is a place for rationality in this field (and in ethics in general), but there is also a limit to what we can achieve by being rational[77].

In the formulation of cognitivism just given, phrases like "having a good reason to accept (believe)" and "justified belief" are included. But as Bergström (1990) has suggested, such phrases are ambiguous (cf. p 91): That a person P's belief that X is justified (that P has good reasons for his belief that X) can either mean (i) that P has a good reason for believing that X, or (ii) that he has a good reason for X. Or alternatively put, a person's belief that X can be justified in the "subjective sense", or in the "objective sense". That P's belief that X is *"subjectively justified"* means (roughly) that it is justified for P (from P's point of view), or rather; that P is justified in holding it (holding on to it, acquiring it)[78]. If P's belief that X is *"objectively justified"*, on the other hand, it is really X that is justified (rather than the belief that X). And that X is justified means that there are good reasons for the validity of X, or alternatively, that there is evidence that supports X (evidence that is available, but not to anyone in particular), and that "everyone" has there-

---

[76]To the extent that justification does not presuppose truth, that is. It can be argued that a claim cannot be (objectively) justified unless it is true, however.

[77]This weak cognitivist view suggests (but does not imply) that some disagreements about what our welfare consists in are "intractable" (they can survive "full discussion and full information", "extensive discussion and awareness of all relevant information"; cf. Harman (1996), pp 10-11), but that there are also disagreements which can be rationally resolved (settled by means of rational argument).

[78]The idea that a person is justified in (has good reasons for) believing that a certain proposition is true is intimately connected to, but not identical with, the idea that it is *rational* for him to hold the belief (given the evidence available to him, and so on). There is (as e.g., Elster (1983) and Nozick (1993) have suggested) more to "theoretical rationality" than having good reasons; for a certain belief to count as rational, it may also be necessary that it has been formed in the right way, e.g., that it has been generated by a reliable procedure (a procedure which is good at generating valid (for example, true) beliefs).

fore (in principle) a good reason to believe that X is true.

When I assume that there are such things as justified value-for-beliefs, it is primarily objective justification that I have in mind. The assumption that our beliefs about what is nonderivatively good (bad, better) for a person can be objectively justified is (in this context) more important than the assumption that these beliefs can be subjectively justified, i.e. the central question is the objective "What counts as a good reason for (or against) a certain value-for-claim?" rather than the subjective "When is it rational for a person to hold on to a certain value-for-belief?" or "How do we determine whether it is rational for a certain person to accept a certain value-for-statement?"[79].

Another source of unclarity is this: It seems that the justification of a claim (or statement) need not be an all-or-nothing matter, i.e. it seems intelligible to say that a certain claim is justified to a certain degree, or that some claims are more justified (more "well-founded", or more rational to accept) than others. In other words, even though there may

---

[79]Here, one might ask whether coherentism and foundationalism should be regarded as answers to the objective "When is the statement S justified?" or "What counts as a good reason for S?", or to the subjective "When is it rational for a person to accept S (to believe that it is true)?". As I see it, foundationalism claims (roughly) that a statement is justified if and only if it is either "self-evident" (e.g., it is obvious that it is true) or it can be "derived" (e.g., induced or deduced) from such self-evident statements, i.e. it is really a conception of objective justification. Coherentism, on the other hand, claims (again roughly) that a person P is justified in believing that X is true if and only if P's belief that X is part of a coherent set of beliefs, i.e. a set where the constituent beliefs are "closely knit together, explaining each other" (Tännsjö (1995), p 574). If this formulation is correct, it seems that coherentism is (at least primarily) a conception of subjective justification. In fact, it seems impossible to conceive of coherentism as a conception of objective justification, and the reason for this is that there is an essential reference to beliefs. However, it seems possible to construct an "intersubjective" version of coherentism, where a reference is made to the beliefs of a certain culture (or the like), rather than to the beliefs of individual persons (How else could a coherentist make sense of science?).

We may also add that epistemological relativism seems plausible if we have subjective justification in mind, but that it can hardly be true if we have objective justification in mind. That is, it does not seem implausible to claim that two inconsistent beliefs (held by different people) can both be "maximally well-founded" (cf. Bergström (1990), pp 117-121), but it seems rather absurd to claim that there are conclusive reasons both for and against a certain statement (the idea that there are "good reasons" both for and against a certain statement is far too weak to qualify as relativism). Part of the reason why this is such an odd view is (I think) that it seems to presuppose ontological relativism (in Bergström's sense) "within the same language community" ("conceptual system", or the like), which is an even odder view.

not be such a thing as a "fully (or maximally) justified" value-for-statement, it is still possible that some value-for-statements are more justified than others, e.g., there may well be such a thing as "the most justified" answers to the questions of prudential value (at least as asked from some specified point of view; cf. e.g., appendix B). So, when I assume that value-for-claims can be justified, what do I mean? Well, I don't mean "fully justified" (that there are conclusive reasons for them, or that they can be proven), but rather that they can be justified to some degree or other.

Another problem is this: Even though it seems reasonable to assume that invalid (e.g., false) claims can be maximally justified in the subjective sense, it can hardly be assumed that invalid claims can be fully justified in the objective sense, i.e. that there can be conclusive (objective) reasons for such claims. So, does it really make sense to say that an invalid evaluative claim be *partly* justified in the objective sense? Well, I think it does, e.g., if there are good reasons both for and against such a claim. That is, if we only have "partial justification" in mind, it seems that we need not assume that all "justifiable" statements are valid.

To sum up, we can formulate the weak cognitivist view as follows: At least some substantive claims of the form "all facts of type X have nonderivative value for all human beings" can be at least partly justified in the objective sense[80].

So, is there any reason to believe that any general theory of prudential value can be *fully* justified in the objective sense, or that any substantive claim of the form "all facts of type X are nonderivatively good (or bad) for everyone" can be justified in this way? Well, there are at least two "sources of doubt": First, it can (as we have already seen) be doubted whether there are any universal prudential values, i.e. whether there are any true general claims about what is good (or bad) for all human beings at all times. If we assume that a claim cannot be fully justified in the objective sense unless it is valid, this would mean that no such general claim can be fully justified in this sense. And second, even if we may assume that there are universal prudential

---

[80]If we remind ourselves that the weak cognitivist view is supposed to hold for the other two kinds of value-for-claims as well (cf. note 75), we get: At least some substantive claims of the forms "the particular fact X is nonderivatively good for a particular person P" and "all facts of a certain type X are nonderivatively good for a particular person P" can be at least partly justified in the objective sense.

values (e.g., pleasure), it can always be doubted whether the relevant general value-for-claims are fully justifiable in the objective sense.

Let us now move on to the next question, viz.

In what general way (or ways) can value-for-claims be justified (or refuted)?

So, on the assumption that at least some claims of the form "all facts of type X are nonderivatively good (or bad) for all human beings" are at least partially justifiable in the objective sense: How can they be justified[81]? Or alternatively, how do we determine whether a general theory of prudential value is well-founded or not? What counts as a good reason for or against such a theory?

This question can be divided into two parts, viz. (i) "How can we justify (or prove) a general value-for-claim? What does a good reason *for* such a claim look like? Is there such a thing as a conclusive reason for such a claim?", and (ii) "How can we criticize (disprove, or refute) a theory of prudential value? What does a good reason against (rational criticism of) such a theory look like? Is there such a thing as a conclusive reason against a general value-for-claim, and if there is, what does it look like?".

Let us now offer some possible answers to these questions. The basic idea here (which is of a coherentist kind) is that the most important requirement that a general theory of value-for must meet is that it is consistent with all plausible (e.g., justified or valid) value-for-judgements, especially those judgements which are of a "more particular" kind, i.e. ultimately, with all plausible particular judgements of the form "The particular fact X is nonderivatively good for P"[82]. This suggests that the major way in which we can justify or refute a general theory of prudential value is by testing it against other (plausible) judgements, ultimately against particular value-for-judgements.

---

[81]As I have already said, the focus will be on the objective question of justification, but on occasion, I will also (when necessary) deal with the subjective question.

[82]The general idea that "more general" statements must be consistent with "more particular" statements can (of course) also be applied to value-for-statements of the "semi-general" kind. If we do this, we get the idea that (semi-general) judgements of the form "All facts of a certain type X are nonderivatively good for a particular person P" cannot be justified unless they are consistent with all plausible particular judgements of the form "The particular fact X is nonderivatively good for P".

The more particular judgements against which a general claim can be tested need not refer to actual situations: the situations referred to may also be imaginary (or purely hypothetical). The best way to make use of imaginary cases is the thought experiment, where hypothetical situations are (normally) *compared* to each other with respect to value-for, and the reason for this is that it allows us to make changes in one variable while we keep the rest constant (it allows use to make use of the locution "other things being equal")[83].

But exactly how do we test the relevant general claims against other plausible judgements? For example, how do we conceive of the relation between the validity of a general statement and the validity of the relevant particular (or semi-general) statements? Here, I want to suggest that (i) and (ii) should be answered as follows:

(i) Here, my (somewhat trivial) suggestion is that there are only two kinds of reasons ("good reasons") that can contribute to the objective plausibility of a general value-for-statement S, viz. (1) that S is compatible with all other plausible (e.g., justified) value-for-statements, especially the particular (and semi-general) ones, and (2) that S can (somehow) be induced from those plausible statements that are of a more particular (or more specific) kind[84]. (If we take the problem of induction into account, this seems to imply that there is no such thing as a conclusive reason *for* a general value-for-statement. This is probably something that a falsificationalist would happily agree with).

In answer to (ii), I suggest (again trivially) that there is only one type of good reason *against* a general value-for-statement S, i.e. that there is only one kind of reason that can "detract from" the (objective) plausibility of such a statement, viz. that S is inconsistent with some

---

[83]A general value-for-judgement can (of course) only be tested against particular value-for-judgements that are "of the same form". For example, comparative thought experiments can only be used to test general comparative value-for-claims, e.g., claims of the form "For all persons P: the more of X (some good-making feature) that a situation contains, the better for P".

[84]This is something that the coherentist and the foundationalist can agree on. The difference between the two views is that while the foundationalist may claim that there are particular statements that are plausible (justified) "in themselves", the coherentist denies this, and claims instead that the reason why the relevant particular statements are justified is (at least in part) that S is justified. (We will return to this idea below).

plausible value-for-statement: or more precisely, that some statement that can be deduced from S is inconsistent with some plausible statement, e.g., a value-for-statement of a more particular kind.

So, does this mean that there is such a thing as a conclusive reason *against* a general value-for-claim? Suppose that there is only one plausible particular statement that is inconsistent with a certain general theory. Does this mean that we should regard the theory as refuted, or should we (instead) "ignore" the particular judgement? What "weight" (or "authority") should we give the particular judgement and what weight should we give the theory?

There are two fundamental views that one might adopt here, views which are best expressed in "subjective terms"[85]. On the first view (which also happens to be my own), we should give a lot of weight to our judgements about particular cases, or more precisely; if there is as much as one plausible (or "considered") particular judgement that is inconsistent with the theory, then the theory must be modified. To gain a better understanding of this thesis, consider the following two points: First, to claim is that our considered judgements about particular cases should be "granted a privileged position" is not to claim that they should be regarded as infallible, or that we never will have a good reason to revise them[86]. And second, it is an open question what it is that makes a judgement considered (but some idea of reflective equilibrium seems plausible here), and how many of our judgements about particular cases that can really be regarded as considered (e.g., it is not enough to be sufficiently convinced). These points suggest that our general theoretical beliefs should be given some weight, but only indirectly, as one of several components in the equilibrium, i.e. it seems that the thesis in question is compatible with coherentism[87]. As Tännsjö (1995) points out,

coherentism /.../ is compatible with a *kind* of foundationalism in et-

---

[85]I hope they can also be expressed in objective terms, but I am not quite sure about this. It all depends on whether coherentism can also be viewed as a conception of objective justification.

[86]One possible reason for this is (of course) that our particular judgements are sometimes theory-laden.

[87]However, it is worth noting the type of coherentism adopted by me is rather "extreme": I tend to give less weight to our theoretical beliefs than most coherentists, including Rawls himself.

hics, a foundationalism according to which certain moral judgements (considered judgements about particular cases, say) are granted a privileged position; these statements are not incorrigible, but, according to other beliefs in the reflective equilibrium, these statements may well have a privileged position (p 574).

The alternative view gives less weight to our considered judgements about particular cases, and more weight to our theoretical beliefs. There are two versions of this idea:

The first version makes the following claim: If there are only a few considered (more) particular judgements that are inconsistent with some general theory, then this does not give us a sufficient reason for rejecting (or modifying) the theory; instead, the relevant particular judgements may "safely be ignored". Since this idea is (on the face of it) contrary to reason, we must ask why some philosophers tend to hold on to it. As far as I can see, this is due to two fundamental convictions (which are, on my view, both implausible, at least in ethics). The first conviction is a too extensive *theoretical ambition* (or a too exaggerated "generalism") in ethics. Only a very theoretically ambitious person might say: "Yes, it is true that my theory has implications which are inconsistent with my considered judgements about particular cases, but I will hold on to it anyway; after all, it is the best general *theory* that we've got, and we *really need* a general ethical theory" (cf. Tännsjö (1993), pp 30-32). The second conviction is the idea that the best (or true) theory must be a *simple* theory (cf. Tännsjö (1995), p 573), who regards lack of simplicity as a considerable intellectual price). On my view, both these convictions are totally unfounded (at least in ethics), e.g., what reason is there (really) to believe that "evaluative reality" (the part of reality that an evaluative theory is about) is *simple*?[88]

The second way in which one might give more weight to general (theoretical) beliefs is this: These beliefs are (so to speak) granted a privileged position *in the reflective equilibrium* (there will be a belief in the equilibrium according to which these beliefs have a privileged position), and they will (because of this) determine (to a high extent) what particular judgements that will be regarded as considered. But if someone

[88]On my view, these convictions also constitute one of the best explanations of why some people are attracted to utilitarianism. For other good explanations of this attraction, see Foot (1985) and Thomson (1992) and (1994).

has a strong conviction that a certain general theory is true (or justified), then this conviction must not (for purposes of justification) have too large an effect on his judgements about particular cases: If someone accepts certain more particular statements as true *because* they are implied by a certain general theory, then coherentism implies that these more particular judgements can not be used to justify the theory.

But as I have already stated, I tend to accept the first view. In short, I think we should (at least in ethics) proceed in "the spirit of inductionism" rather than in "the spirit of deductionism".

All this gives rise to the following question: If general value-for-claims are (at least in part) justified in virtue of being consistent with more particular judgements which are themselves justified, how can *these* judgements be justified? That is, how (in what general way) can claims of the forms "all facts of type X are good for a particular person P" and "the particular fact X is good for a particular person P" be justified? What kinds of reasons can be given for and against such claims?

If we look at how the semi-general claims can be justified, there are two possibilities here, viz. these claims can perhaps (1) be justified in an "inductive" way, but they may also be (2) justified in a "deductive" way.

As far as the "inductive" type of justification is concerned, nothing really needs to be added to what has been said about general value-for-statements above. In short, there are two kinds of good reasons *for* such a semi-general statement S, viz. (a) that S is consistent with all other plausible value-for-statements, especially the particular ones, and (b) that S can (somehow) be induced from the relevant particular statements; and there is only one type of good reason *against* a semi-general value-for-statement S, viz. that some value-for-statement that can be deduced from S is inconsistent with some plausible statement.

However, there is also the possibility that semi-general value-for-statements can be justified in a "deductive" way, i.e. in terms of general value-for-statements. It might, for example, be claimed that it is non-derivatively good for a certain person to feel pleasure because he is a certain type of creature (e.g., a sentient being, or a human being), and because feeling pleasure has nonderivative value for all creatures of this type. It is worth noting that this deductivist idea is not inconsistent with the "inductivist" idea that we should give more "epistemological weight" to our more particular judgements than to our more general

judgements; it may (after all) be "partially valid" for coherentist reasons. It seems inconsistent with the strong "inductionism" expressed above, however (this is not changed by the fact that my "inductionism" is "falsificationalist" rather than "verificationalist" in spirit), and this suggests that the deductivist idea is of little or no help in this context.

As far as the justification of particular value-for-statements[89] is concerned, it seems plausible to adopt the following generalist view: If a certain particular fact has nonderivative value for a certain person, then the reason for this is (at least in part) that the fact in question has certain good-making features (and lacks certain bad-making features). But this would mean that all facts of the same type (i.e. facts that have the same features) are also good for the person. In short, whenever a particular fact has nonderivative value for a particular person, then this is so because it is of a certain type, and because all facts of this type have nonderivative value for the person in question (This is the idea of supervenience which is discussed more in detail in appendix B).

Now, this view may seem inconsistent with the "inductionist" view that a semi-general value-for-statement can be partly justified in virtue of being consistent with all plausible particular value-for-statements, but the inconsistency is (I think) only apparent. The generalist view above does not imply that particular value-for-statements are justified in virtue of being deducible from plausible semi-general statements, i.e. it is not really a "deductionist" view. But still, the generalist view has (I think) certain "justificatory" implications, viz. it suggests that the value-for of particular facts should (in this context) not be regarded in isolation (we will soon return to this idea).

This suggests that we should adopt the following "coherentist" (and "generalist") view: A particular value-for-statement S is at least partially justified (in the objective sense) if it is consistent with all other plausible particular statements, where S is consistent with another particular statement S' if and only if S' can be deduced from some plausible semi-general statement that corresponds to S[90]. (And we should also add

---

[89]It is worth noting that these statements will be ignored in the rest of this book. My only reason for discussing them at all is that they have been granted the most privileged epistemological position.

[90]That is, the idea is that a particular statement S can only be consistent with other particular statements *via* some semi-general statement to which it corresponds: if a particular fact X is nonderivatively good for P, then there exists some description of X such that all other particular facts that fall under the same

that there is no such thing as a self-evident particular value-for-belief).

All this suggests that there is no such thing as a conclusive reason for or against a certain particular value-for-statement (which is not to deny that the validity of such statements are sometimes "beyond reasonable doubt"). It also seems to have certain implications on the issue of "subjective justification", e.g., it seems to imply that there is little or nothing one can do to resolve disagreements about the value-for of particular facts in a rational way (unless we conceive of rhetoric tricks as rational argument, that is). If this is correct, we must also (as a consequence) accept that those "general (e.g., theoretical) disagreements" which are based on such "particular disagreements" are also intractable, and that it is in many cases not possible to convince "every rational person" (by means of rational argument) that a certain general value-for-claim is valid or invalid. In fact, it seems likely that we can only convince someone who shares the relevant intuitions about particular cases. For example, suppose I argue against a certain general theory by trying to show that it has implausible implications in more particular cases, e.g., suppose I say: "Hedonism implies that life in the experience machine (cf. section 3.2 below) can be good, but such a life can never be good, and therefore, hedonism is implausible". What do I do if the opponent thinks that a life in the experience machine may well be a good life[91]? If

---

description (that are of the same type) are nonderivatively good for P too. Or in other terms: For every plausible particular statement of the form "this is good for P", there are a number of a corresponding plausible semi-general value-for-statements on different levels of generality, e.g., "it is good for P that he has a pleasant musical experience", "it is good for P that he has a pleasant aesthetic experience", or "it is good for P that he has a pleasant experience". And the corresponding (plausible) semi-general statement that is of most interest in this context is (of course) the most general one: "it is good for P that he has a pleasant experience" is (if valid) of more interest than "it is good for P that he has a pleasant musical experience", and so on.

That a certain particular statement is consistent with other particular statements in this way (*via* some semi-general statement) is probably the strongest type of reason one can find for regarding it as valid. To say that a particular statement is justified if it can be deduced from some justified general statement is (as we have seen) not very helpful in this context. And it is also doubtful whether there is such a thing as an evaluative counterpart to "good observational conditions"; we could (of course) always refer to the good (e.g., mentally healthy) human being, but this would (it seems) only give rise to new difficult problems.

[91]It is also possible that the opponent does not accept that the theory has a certain implication, e.g., like those hedonists (yes, I have met some) who refuses to accept that the theory implies that a pleasant life in the experience machine is a good life. How does one respond to someone who refuses to accept one's "data" as data?

we want to resolve our disagreement in a rational way, how do we go on from here? For example, can we resolve our disagreement by moving on to another "level of discourse", e.g., by starting to talk about the nature of our disagreement? I think not. On my view, there is really nothing we can do here.

This pessimistic view concerning the "foundations of ethics" gives rise to at least two questions, viz. (i) if I accept the idea that there is no such thing as an evaluative counterpart to "good observational conditions" in science, why do I give such epistemological weight to our judgements about particular cases?, and (ii) why do I hold on to cognitivism at all; why don't I take one step further and become a full-fledged scepticist?

(i) My reason for giving more epistemological weight to our judgements about particular cases than to our more general judgements is simple: It is true that we may well be mistaken about what particular facts that are nonderivatively good for us, but it is more likely that we are mistaken in our more general beliefs, viz. because the latter beliefs contain much more information. Just as the belief that all allergies are psychosomatic is more likely to be false than the belief that this particular allergy is caused by psychological factors, so the value-for-belief that it is nonderivatively good for everyone to feel pleasure is much more likely to be false than my belief that it is nonderivatively good for me to feel whatever I am feeling right now. So, even though our judgements about particular cases may be mistaken; they are at least more certain than the alternatives.

(ii) So, in the light of my "foundational pessimism", what reasons do I have for rejecting full-fledged scepticism, i.e. the idea that there is no such thing as an objectively good reason in this area, and that our value-for-beliefs can be neither justified nor refuted? Well, my only reason for rejecting scepticism about value-for is the idea that particular value-for-statements can (for "coherentist reasons") be at least partially justified. If this is not a good reason, then I am (it seems) forced to accept scepticism.

(And how does one respond to someone who thinks we should consider the fact that the experience machine may break down??!).

Chapter Two

A formulation of the hedonistic theory (and its different versions)

The first traditional conception of prudential value which we will look at is the hedonistic theory. The discussion of hedonism will have the following structure. In this chapter, I will try to give a formulation of the theory (and its different versions) that is as precise as possible. In connection with this, several topics will be discussed, e.g., "What is the nature of pleasantness (and unpleasantness)?" and "To what is pleasantness primarily attributed, to 'whole lives of experience at certain times' or to 'experiences' (in the plural)?". The central questions in chapter 3 are questions of plausibility. By looking at arguments that can be given for and against the theory, I will try to find out whether the hedonistic theory is a plausible theory of prudential value or not, and which version of the theory that is most plausible.

To recapitulate how the pure version of the hedonistic theory was characterized in chapter 1, here is a brief summary:

(H1) The Experience Requirement: The only facts that can have nonderivative value for a person at a certain time are facts about his or her own experience at that time.

(H2) More specifically, the only thing that is nonderivatively good for a person is to have pleasant experiences, and the only thing that is nonderivatively bad for a person is to have unpleasant experiences.

(H3) The Thesis of Unrestrictedness: All pleasant experiences are nonderivatively good for the experiencing subject, and all unpleasant experiences are nonderivatively bad, regardless of what other properties these experiences have.

(H4) If we ignore the issue of duration, then every good experience is good in virtue of its pleasantness only, and every bad experience is

bad in virtue of its unpleasantness only.

(H5) The "intensity-orientation" and the idea of proportionality: To the extent that it is possible to determine how valuable it is for a certain person to have a certain experience: The value of an experience for the person who has it is (if we ignore the issue of duration) a function of one thing only, viz. how pleasant or unpleasant it is. Or more specifically: The positive (or negative) value that a pleasant (or unpleasant) experience has for the experiencing subject is proportional to how pleasant (or unpleasant) the experience is. This means (among other things) that the marginal value of pleasure (and displeasure) is not diminishing.

(H1)-(H5) are claims about the value-for of local situations (or "atomic facts"). (H6), on the other hand, is a claim about the value-for of whole lives-at-certain-times, situations the value-for of which cannot be determined directly.

(H6) The final value that a certain life-at-a-certain-time has for the person who is living it is a function of how much pleasure and how much suffering this life contains. The more pleasure it contains, the better, and the more suffering it contains, the worse.

As it stands, this claim is far too imprecise. So, how would a pure hedonist make it more precise? Well, this is the position that he would most likely adopt:

(H6:a) First, he would (of course) assume that *evaluative atomism* is true, i.e. (in this particular case) that the final value of a person's existence at a certain time has for this person is a function (e.g., a sum) of the (positive and negative) nonderivative values-for of its constituent parts[1].

So, how should this function be characterized, i.e. how does the hedonist tackle the problem of synchronic aggregation? Well, as a first approximation, we know that the hedonist would adopt the following

---

[1]Where "parts" should (I think) be understood in a broad sense, e.g., as "aspects" or "features". However, in a hedonistic context, it doesn't really matter whether we use the term in a strict sense or in a broader sense. The reason for this is that on the hedonistic view, the only relevant feature of a complex situation is its experiential "content", something that can be referred to as "parts" in the strict sense.

rudimentary theory of aggregation: The more pleasure and the less suffering that a certain life contains, the better it is for the person who is living this life. But how exactly do we determine how much pleasure or suffering a life contains? And how do we deal with the cases where one possible life contains both more pleasure and more suffering than another possible life? The hedonistic answer to these questions can be divided into two parts:

(H6:b) This is the procedure that should (in principle, and if possible) be adopted in order to determine how much pleasure and how much suffering a life contains: We assign a positive number to each pleasant experience, and a negative number to each unpleasant experience, that the life contains. (How great these numbers are depends on how pleasant or unpleasant the experiences are). We then add all the positive numbers we have assigned to the pleasant experiences, and all the negative numbers we have assigned to the sufferings. If we do this, we get two sums, viz. the positive sum Σ (P), and the negative sum Σ (S). We can now say that the greater the sum Σ (P) is, the more pleasure a life contains, and the more negative the sum Σ (S) is, the higher the amount of suffering[2].

(H6:c) Once our hedonist has access to the two sums Σ (P) and Σ (S), he can make the idea that "the more pleasure and the less suffering that a certain life contains, the better for the person who is living this life" more precise, viz. in the following way: The greater Σ (P) is, the better, and the smaller Σ (S) is, the worse.

But how exactly should we understand the idea that the value-for of a life is a function of the two sums Σ (P) and Σ (S)? That is, how should we (according to the hedonist) calculate the final value-for of a life from its Σ (P) and Σ (S)? There are two different ways in which a pure hedonist might answer this question, viz. he can either appeal to (i) to the difference thesis, or to (ii) some kind of ratio thesis.

(i) *The Difference Thesis* is (in the case of lives-at-certain-times) the idea that a life L1 is finally better (more pleasant) than another life L2 if

[2]It is important to point out that for this type of aggregation to be at all possible, two rather dubious assumptions has to be made, viz. (i) both pleasantness and unpleasantness are measurable on ratio scales, and (ii) both pleasantness and unpleasantness satisfy "the criterion of additivity".

and only if L1 contains more "net pleasure" than L2, i.e. iff "the total net sum" [Σ (P1) + Σ (S1)] is bigger than the sum [Σ (P2) + Σ (S2)]. Or alternatively put, L1 is better than L2 if and only if the difference [Σ (P1) - | Σ (S1) |] is bigger than the difference [Σ (P2) - | Σ (S2) |].

(ii) *The Ratio Thesis*, on the other hand, is (roughly) the claim that a life L1 is better than another life L2 if and only if L1 has a better "balance of pleasure over pain" than L2. But how exactly should this idea be understood? Well, the idea is most certainly *not* that L1 is better than L2 if and only if the ratio Σ (P1)/ | Σ (S1) | is bigger than the ratio Σ (P2)/ | Σ (S2) | (this would have the absurd implication that one tiny pleasure is enough to make a life free from suffering "infinitely good"). This is one of several possible interpretations that is far more plausible: L1 is better than L2 iff the ratio Σ (P1)/[Σ (P1) + | Σ (S1) |] is bigger than the ratio Σ (P2)/[Σ (P2) + | Σ (S2) |][3].

This is how the pure hedonist would make (H6) more precise. Let us now turn our attention to (H1)-(H5). As these claims stand, they are not sufficiently clear and precise for our purposes, i.e. in order to find out how plausible the different versions of hedonism are, we need to formulate the claims in a more precise way. So, what is it about these claims that is in need of clarification? What is it we need to know in order to arrive at a sufficiently precise formulation of pure hedonism? On my view, there are two "sources of unclarity", viz. the following ones:

(1) When the hedonist claims that the value of an experience (for the person who has it) is dependent on how pleasant or unpleasant it is, it is not entirely clear what the terms "pleasant" and "unpleasant" mean,

---

[3]Here, it is important to notice that for these "operations" are not possible unless two very strong conditions are (in addition to (i) and (ii) in note 2, that is) met, viz. the following ones: (iii) Pleasantness and unpleasantness are measurable on the same ratio scale, something which (among other things) presupposes that the positive values of our pleasures can be fully compared to the negative values of our sufferings: if P is a pleasure and S is a suffering, then either the negative value of S is "smaller than", or it is "as big as", or it is "larger than" the positive value of P. (iv) It is not just that we can obtain the two sums Σ (P) and Σ (S), we can also (once we have obtained them) perform arithmetic operations on these sums. That is, the two sums can be added, divided by each other, and so on. Or alternatively put, it is not just meaningful to add the positive (or negative) numbers assigned to a person's pleasures (or sufferings); it also makes sense to add (etc.) the positive numbers assigned to his pleasures and the negative numbers assigned to his sufferings to each other.

and what they refer to. So the first thing we need to know is how the terms "pleasantness" and "unpleasantness" should be interpreted in this context, i.e. how these key terms are actually used by the hedonist when he states his claims. We also want to know which of the possible interpretations of the terms that makes the hedonistic theory most plausible, or alternatively put, which of the possible senses of the terms "pleasant" and "unpleasant" that are most ethically relevant, but with the following "proviso"; this interpretation must not deviate too much from the ordinary meaning (or meanings) of the terms.

These questions need not be formulated in semantic terms, however; they can also be formulated in terms of conceptions of pleasantness and unpleasantness. A *conception of pleasantness* is a theory that purports to give answers to certain central questions about pleasantness, e.g., questions like the following ones: What is it for someone to have a pleasant experience? What is it that makes an experience pleasant: what does the pleasantness of an experience consist in? Do all experiences which are properly conceived of as pleasant have anything in common, some property (intrinsic or relational) in virtue of which they are pleasant, or is the class of pleasant experience a "heterogeneous class"? That is, are all pleasant experiences of the same *genus* (in the traditional sense of the term) or not? And what fundamental forms (e.g., *species*) of pleasantness are there?

So, what conceptions of the pleasant and the unpleasant does different hedonists have in mind? And what conception of pleasantness and unpleasantness makes (if combined with the hedonistic theory) the theory most plausible, i.e. what conception of pleasantness and unpleasantness is (so to speak) most ethically relevant?[4]

---

[4]It is also worth mentioning what we are *not* really interested in in this context. We are not looking for the true meaning (nominal definition) of the terms "pleasant" and "unpleasant", a meaning which captures the real nature (or essence) of pleasantness and unpleasantness. Even if there were such a thing as the nature of the pleasant and the unpleasant (which is doubtful), and even if knowledge about this "thing" would (so to speak) give us "the true meaning of hedonism", we could still ignore it. What we want to know is not what pleasure (displeasure) *really is*, but what the hedonist *really means*, and above all, what he *should mean*. Neither are we particularly interested in what the terms "pleasant" and "unpleasant" mean in ordinary speech, nor in how we should conceive of these terms "in a defensible psychological conceptual scheme", i.e. in what conceptions of the pleasant and the unpleasant that are "most useful for psychological theory" (cf. Brandt (1979), p 35).

(2) The hedonist conceives of pleasantness and unpleasantness as features of experience, i.e. as something that is properly attributed to experience. But when pleasantness or unpleasantness are attributed to "experience", to what exactly are they attributed? Is pleasantness something that is primarily attributed to whole lives of experience at certain times, or is it something that is primarily attributed to experiences (in the plural, in the sense in which a person can have several experiences going on at the same time)[5]? Is pleasantness something that *can* be attributed to whole experiential lives at certain times, and if it can, is it something that can be *directly* attributed to whole experiential lives? And if it is more plausible to conceive of pleasantness as a feature of experiences (in the plural), then how should such experiences be individuated "in the synchronic", i.e. how do we separate two simultaneous experiences from each other?

These are the questions we have to answer in order to arrive at a precise enough formulation of pure hedonism[6]. Let us start with the second one, i.e. the question of what pleasantness and unpleasantness are (primarily) attributed to.

2.1. To what exactly is pleasantness and unpleasantness primarily attributed?

Some preliminaries: How we attribute pleasantness and unpleasantness in ordinary speech

In everyday speech, pleasantness is attributed to many different kinds of things, e.g., to "sniffing a perfume, eating a steak, dozing in a deckchair, playing tennis, a general euphoria or sense of well-being or content, announcing to someone else that he is receiving an award, listening to a symphony, watching a football game, making preparations

---

[5]This is how I expressed myself when formulating (H1)-(H5).

[6]But what if the value of an experience for the person who has it is not just dependent on how pleasant or unpleasant it is, but also on its duration (on how long it lasts)? This would give rise to yet another "source of unclarity", viz. the following one: How exactly should the term "duration" should be understood here? Should it be understood in the objective sense or in the subjective sense? Is it objective time (clock time) or subjective time (felt time) that matters? There will be a detailed (and rather speculative) discussion of these questions in appendix E.

for a trip, day-dreaming about some hoped-for event, and so on" (Brandt (1979), p 35), or to "satisfying an intense thirst or lust, listening to music, solving an intellectual problem, reading a tragedy, and knowing that one's child is happy" (Parfit (1984), p 493).

This list indicates that in ordinary speech, pleasantness is not only attributed to experiences (like euphoria or day-dreaming), but also to "doings" or *activities* (where the term "activities" should be understood in a broad, "Aristotelian", sense, i.e. as including the activities of the mind), like eating, playing tennis, or solving a problem[7]. So, in virtue of what do we attribute pleasantness to activities? What is it to take pleasure in an activity, and what is it that makes a pleasant activity pleasant?

On my view, there are three kinds of reasons for attributing pleasantness to a certain person P's engaging in a certain activity, viz.
(i) it (or the experience of it, which I regard as included in the activity) causes that P has a pleasant experience; or
(ii) it is the intentional object of some pleasant experience that P has, e.g., P is happy with engaging in it; or
(iii) P enjoys to be engaged in the activity, where the term "enjoys" is to be understood as follows:

> Someone *enjoys* an activity to the extent he engages in the activity because of its own intrinsic properties, not simply because of what it leads to or produces later. Its intrinsic properties are not limited to felt qualities, though; this leaves open the possibility that something is enjoyed yet not pleasurable (Nozick (1989), p 104)[8].

That is, an activity is not enjoyed (in this sense) if the agent desires to engage in it because of the experiences it gives rise to. Rather, that an agent enjoys an activity means that he desires to engage in it "in and for itself", i.e. that he has an *intrinsic desire* to engage in the activity,

---

[7]On this view, it is not appropriate to attribute pleasantness to things that are neither experiences nor activities, e.g., to external events. Now, it may seem as if pleasantness might sometimes be correctly attributed to "things" like games or pieces of music, but this is not really so. What we attribute pleasantness to in cases like these are (rather) things like *watching* a game, or *listening* to music.
[8]Where it is not entirely clear what an intrinsic property of an activity is. For example, if someone performs a certain action because he believes it is his duty, is this action enjoyed in Nozick's sense?

supposedly *when* he engages in it[9].

To conclude: If pleasantness is (correctly) attributed to something that is not an experience, then (a) this something is an activity (in the wide sense), and (b) the reason why the activity is properly regarded as pleasant is either that it is (in some relevant way) connected to some pleasant experience that the agent has, or that it is enjoyed or liked. That is, we are left with two things, viz. pleasant experience and enjoyment. Enjoyment is of little or no interest to most hedonists, however[10], so from now on, we will restrict our attention to pleasant experience.

In ordinary speech, we rarely attribute pleasantness and unpleasantness to whole experiential lives-at-certain-times. Instead, pleasantness and unpleasantness are normally attributed to certain parts of such "synchronic wholes", viz. to different kinds of *experiences* (in the plural). More specifically, the kinds of experiences to which pleasantness and unpleasantness are most often (and most naturally) attributed are *feelings* (in the broadest sense of the term), viz. *bodily sensations*, *emotions*, and *moods*.

---

[9]It is worth pointing out that Griffin's (1986) use of the term "enjoyment" is similar to Nozick's. This is what Griffin writes about enjoyment: "There is a cluster of terms which even in their everyday use seem to fall conveniently between mental states and fulfilment of desires: namely, enjoying or liking things, finding them pleasing or satisfying or fulfilling, being pleased or happy with them. Let us use the term *enjoyment* to cover them all. *Enjoyment* /.../ is not anything so narrow as experiencing a single state or one of a range of states /.../ [and] it is nothing so broad as having merely desires /.../ fulfilled. Also, let us allow that people *enjoy* things other than states of mind [e.g., helping others or advance knowledge]. /.../ But let us put a limit to the range by requiring that all the objects of *enjoyment* fall within our experience" (p 18).

[10]Or am I wrong about this? I can't help feeling that the exclusion of enjoyment is (somehow) against "the true spirit of hedonism". Anyway, it is worth considering what an *"enjoyment-hedonistic" theory* would look like: It would claim that the only thing that is nonderivatively good for a person is that he enjoys doing what he is doing, and the only thing that is nonderivatively bad for a person is to be engaged in activities that he does not enjoy at all, or to be engaged in activities which he positively dislikes to be engaged in. It would claim that the more enjoyed activity and the less disliked activity a life contains, the better it is for the person who is living the life. And it would claim that how good it is for a person to engage in an activity depends on how much he enjoys it, i.e. to what extent he engages in it because of its own intrinsic properties. On my view, this theory is not bad at all.

Pleasant and unpleasant sensations

There are at least two things that all sensations have in common: First, they are not directed towards intentional objects, and second, every sensation is associated with some sense organ (or type of receptor). This last fact provides us with the most natural way of classifying sensations, viz. on basis of what kind of sense (or receptor, or neural pathway) that the sensation is associated with. In this way, we distinguish between visual sensations, sensations of taste or smell, tactile sensations, sensations of cold, sensations of pain, and so on.

It seems that the kinds of sensations to which we normally attribute pleasantness and unpleasantness are the most "physical" ones, i.e. the sensations which are associated with the proximal senses, like tactile sensations, sensations of smell, sensations of taste, sensations of pain, and so on. That is, visual and auditory sensations (e.g., seeing a patch of red) are (unlike visual and auditory perceptions, which also have cognitive components) not often conceived of as pleasant or unpleasant.

Examples of sensations that are normally (but not always) pleasant to have are the taste caused by a good wine; the bodily feelings that are caused by things like getting a massage, relaxing, or taking a hot bath; the sensations of cold in the throat that one gets from drinking something cold when thirsty; the sensations of touch that are caused by kissing and caressing a person that one finds attractive, or by being kissed or caressed by such a person; and so on. Examples of sensations that are normally considered unpleasant are sensations of pain, itchings, or the sensations that are associated with things like the following; receiving an electric shock, having to stay awake when sleepy, being unable to breathe (or to have one's breathing restricted), being (bodily) tense, being too cold or too hot, tasting something really sour, or smelling something rotten.

Pleasant and unpleasant emotions

Emotions are intentional mental states or events, i.e. they are always directed towards (intentional) objects. Every emotion has (in virtue of its intentionality) some cognitive content, it is in part constituted by "cognitions" (in a wide sense of the term) about its object. These cognitions are (it seems) of two different kinds: they are either beliefs

about the object[11], or they are "pro- or con-attitudes" towards the object. That is, when someone is emoting, he believes something about the object, and he is (perhaps as a result of having these beliefs) either "for" or "against" the object, e.g., he evaluates it in a positive or negative way, he likes or dislikes it, or the like. Emotions are not purely intentional states (like thoughts or fantasies), however; they also have sensory content, or more specifically, they are (on my view) partly constituted by bodily sensations (or felt qualities of a "bodily" type)[12].

Examples of emotions that are normally (but not always) pleasant to have are happiness[13], gladness, joy, infatuation, pride, and hope, and examples of emotions that are normally unpleasant are grief, anger, fear, worry, despair, guilt, shame, hatred, and unhappiness. Most emotions that are considered pleasant are also "positive" (they are, in part, constituted by a positive evaluation of its object), and most unpleasant emotions are "negative", but there is (as far as I can see) no necessary connection between pleasantness and "positivity", or between unpleasantness and "negativity", e.g., it may well be the case that an emotion is both positive and unpleasant (as in the case of intense longing).

Pleasant and unpleasant moods

The third type of experience to which pleasantness and unpleasantness are most typically attributed is the moods. According to Nozick (1989), a mood is a "tendency to make certain types of evaluations, to focus

---

[11]Where "perceiving something in a certain way" might be regarded as a form of believing. That is, if I perceive a bull as dangerous (if the bull appears dangerous to me), then it is (in this "emotional" context) appropriate to ascribe to me a belief that the bull is dangerous.

[12]All this is in line with Nozick's (1989) theory of emotion. On his view (cf. pp 87-89), emotions have a common structure, consisting of three components: (i) a belief that something is or is not the case, (ii) a positive or negative evaluation, and (iii) a feeling, sensation, or "inner experience".

[13]Emotional happiness is probably the most pleasant emotion there is, especially the type of happiness that Nozick (1989) calls *"feeling that your life is good now"*. This is how he describes it: "Recall those particular moments when you thought and felt, blissfully, that there was nothing else you wanted, your life was good then. /.../ What marks these times is their completeness. There is something you have that you want, and no other wants come crowding in; there is nothing else that you think of wanting right then. /.../ [I]n the moments I am describing, these other desires /.../ simply are not operating. They are not felt, they are not lurking at the margins to enter. There is no additional thing you want right then, nothing feels lacking, your satisfaction is complete. The feeling that accompanies this is intense joy" (pp 108-109).

upon facts that can be evaluated in that way, and to have the ensuing feelings" (p 114). I agree with Nozick that moods are, to a considerable extent, dispositions (or "filters"). However, on my view, a mood is more than just a set of dispositions; it is also (at least in most cases) constituted by (i) a certain "outlook" (on the world), and (ii) a diffuse (perhaps global) feeling (cf. Brülde (1992), pp 98-99).

(i) It seems that every mood is (at least in part) constituted by a certain way of "perceiving" the world. To be in different moods is to see the world in "different lights", or even (in some sense) to "live in different worlds". (That is, moods are not intentional states, they are not directed towards specific objects). To describe the "light" in which we see the world, we often use colour metaphors (like in "all is grey today"), or we talk in terms of darkness and light (like in "today, I see everything from the bright side"). Being in a mood is like wearing a pair of glasses, and we are always wearing some pair or other, i.e. we are (as Heidegger has pointed out) always in some particular mood or other.

(ii) To be in a certain mood is (at least in most cases) to feel in a certain way: Most (perhaps all) moods are partly constituted by feelings. These constitutive feelings are, I suggest, bodily feelings, but most of them do not have a distinct location; instead, they are rather "diffuse", and they are typically attributed to the body as a whole. Examples of such "global states of feeling" are; to feel calm, empty, tense, relaxed, or low; to feel "nothing" (to be numb); to feel at ease (to be comfortable); to feel tired, sleepy, awake, alive, or energetic. To describe these feelings (and the mood-states which contain them) we use (more or less metaphorically) term pairs like "full vs. empty", "awake vs. asleep", and "alive vs. dead". However, the most common metaphor in this area is probably "high vs. low": we are (in a metaphorical sense) always on some "energy level" or other, e.g., when we feel depressed, we are in a low mood.

Examples of moods that are normally (perhaps always) pleasant to be in are "mood-happiness" (happiness *qua* mood), "mood-joy", and harmony, and examples of moods that are typically unpleasant to be in are depression, anxiety, boredom, "mood-sadness", and "mood-anger" (to be irritated in general).

The global character of the moods makes them very important from a hedonistic point of view. How pleasant or unpleasant a person's total

state of mind is at a certain moment is (to a considerable extent) dependent on what mood he is in, e.g., listening to good music or drinking a good wine is not particularly pleasant for a depressed person.

So, are sensations, emotions, and moods the only kinds of experiences to which we attribute pleasantness and unpleasantness, or do we also attribute these properties to other kinds of mental states and events, e.g., to thoughts, fantasies, or (visual or auditory) perceptions? Well, we most certainly do if these states are accompanied by some pleasant feeling. However, I don't really know whether we attribute pleasantness to fantasies or perceptions "regarded in isolation". In any case, pleasantness and unpleasantness are typically and normally attributed to experiences that are (at least in part) constituted by some bodily sensation, i.e. to experiences that have "felt qualities". (And the only kinds of experiences that are of this type are bodily sensations, emotions, moods, or more complex experiences which contain sensations, emotions, or moods as elements).

Pleasant experiences vs. pleasant "experience"

Now that the "preliminaries" are over, let us turn to question (2) on p 73. Is pleasantness primarily attributed to total experiential states at certain times (as "the total view" claims), or is it primarily attributed to experiences (in the plural), i.e. to certain parts of our total conscious mental states (as "the partial view" claims)?

The issue is not whether it is possible (and meaningful) to attribute pleasantness to whole experiential lives-at-certain-times and/or to experiences in the plural. To adopt the partial view is not to deny that pleasantness can be meaningfully attributed to total experiential states, and to adopt the total view is not to deny that we can attribute pleasantness to "experiences" (assuming that there is some way in which experiences can be separated from each other).

So, how exactly do the two views differ from each other? Well, "the total view" claims that pleasantness is *primarily* ascribed to total experiential states. As I see it, this claim can be divided into two parts, viz. (i) the idea that pleasantness can be *directly* attributed to total conscious mental states (at certain times), and (ii) the holistic idea that in order to determine how pleasant or unpleasant a certain separate experience is,

we must not regard it in isolation; instead, we should regard it in relation to the organic whole of which it is a part, and the reason for this is that the pleasantness of the part is somehow conceived of as dependent on the pleasantness of the whole.

"The partial view" rejects the first of these claims, e.g., it denies that pleasantness can be *directly* attributed to a person's total experiential state at a certain time. Instead, it claims that pleasantness can only be directly attributed to separate experiences. This view implies (I think) that the only way in which we can determine the pleasantness or unpleasantness of an experiential whole is by "deriving" it from the pleasantness (unpleasantness) of its constituent parts. This means that if the holist is right and the atomist wrong (if the pleasantness or unpleasantness of a total experiential state is *not* a function of how pleasant or unpleasant its constituent pleasures and displeasures are), then it is not really possible to determine just how pleasant or unpleasant a certain experiential whole is.

So, which view is most plausible, the total view or the partial view? Let us first note that the total view has certain features that makes it attractive to the hedonist, viz. *if* it is correct, then the hedonist can ignore both the problem of synchronic aggregation and the problem of how concrete experiences can (and should) be separated from each other (how a person's total experiential state at a certain time is to be divided into experiences). However, the presence of these features does not give us any reason to regard the view as correct. The same thing holds for the perhaps most attractive feature of the partial view, viz. that it is more in line with the way we speak, think, and act in our everyday lives. In ordinary speech, pleasantness is normally (perhaps always) attributed to separate experiences, like sensations, emotions, or moods. And most of the time, we also seem to think (and deliberate) in terms of separate experiences rather than in terms of experiential totalities, e.g., as in "if you do this or that, you will have a great experience".

What we need to know in order to determine which view is most plausible is whether pleasantness can be directly attributed to a person's total experiential state or not. Now, on my view, this is not possible. This does not mean that I accept atomism, however; the claim is rather (i) that it is almost always possible (and often necessary) to divide a person's total experiential state (at a certain time) into parts, some of which are more or less pleasant while other parts are more or

less unpleasant, and (ii) that the pleasantness or unpleasantness of a total conscious state is (in part) dependent on how pleasant or unpleasant its constituent experiences are (i.e. there is *some* truth in atomism).

Now, it might seem that it can often be directly determined how pleasant or unpleasant a person's total experiential state is (e.g., by the person himself), but this is not really true. On my view, it is often possible for a person to determine how pleasant or unpleasant his mood is, but a person's mood state (a relatively homogenous state which can remain the same for days or more) should not be confused with his total experiential state (a heterogeneous and complex state which changes from moment to moment).

As far as "the problem of individuation" is concerned, I don't think it constitutes much of a threat to the partial view. First, it is not very difficult to divide our total sensory experience into parts (i.e. on basis of sense modality), and second, even though there is no non-arbitrary way to divide a person's "emotional life" into parts, there is a way in which this can be done, viz. the one that is provided by our ordinary language.

If I am right about all this, we can safely stick to ordinary speech. We can continue to talk in terms of experiences (in the plural), and we can be "sure" that it makes sense to attribute pleasantness to sensations, emotions, moods, and so on. We can discuss the pros and cons of hedonism in terms of experiences, and we can ask what the pleasantness of a particular (separate) experience consists in.

Let us now turn to question (1) on pp 71-72, i.e. the question of what the terms "pleasant" and "unpleasant" are (and should be) referring to in this context.

2.2. What do the terms "pleasantness" and "unpleasantness" refer to? Different conceptions of pleasantness and unpleasantness

Before we look at the most common (and most plausible) conceptions of pleasantness and unpleasantness, let us first formulate some platitudes which every plausible conception must accept. The platitudes I have in mind can all be regarded as variations on a single theme, viz. the idea that it is of great importance to make a distinction between pleasure

and displeasure (suffering), on the one hand, and pain (i.e. physical pain) and "plain"[14] (the positive counterpart to pain, assuming that there is such a thing), on the other. In the following, I will restrict my attention to the distinction between suffering (or displeasure) and pain (it is, after all, doubtful whether there are such things as "plains", or "sensations of pleasure").

(i) The term "pain" refers to one or several kinds of mental object (event or state), viz. to one or several kinds of sensation. That is, to be in pain is to have a special, distinct kind of sensation, a sensation which is properly described as a "sensation of pain", rather than as a "painful sensation". The term "displeasure", on the other hand, does not refer to any special kind of mental object (like "sensations of displeasure"), but to those objects which have the property (quality or relational property) unpleasantness. To maintain awareness of the fact that displeasure and suffering is really a matter of properties of experiences, I suggest that we replace the nouns "displeasure" and "suffering" with the corresponding adjectives and/or adverbs, i.e. "unpleasant" (but "unpleasantness" will also do). That is, we should talk in terms of unpleasant experience rather than in terms of displeasure or suffering.

There are a number of other ways in which unpleasant experience can (and should) be distinguished from the sensation of pain:

(ii) Not all unpleasant experiences belong to the category of sensation; e.g., we also attribute unpleasantness to moods and emotions. And all moods and emotions which are properly classified as unpleasant are not unpleasant because they contain sensations of physical pain (or even unpleasant sensations) as components.

(iii) But even if we restrict ourselves to the case of sensation, the distinction still stands. To see why unpleasant sensations have to be distinguished from sensations of pain, we only have to consider first, that it is not necessarily unpleasant to be in pain (it might even be pleasant), and second, that every unpleasant sensation is not a sensation of pain, e.g., as in the case of certain itchings and muscular tensions.

Once the distinction between the unpleasant and the painful is made, it is obvious that what the sensible hedonist is primarily concerned with is the former rather than the latter. More specifically, the hedonist would claim: (a) that it is not bad for a person to have a sensation of

[14]This is my translation of Furberg's (1993) "närtor" (as opposed to "smärtor").

pain unless it is also unpleasant, and (b) that it is bad for a person to have an unpleasant sensation, even if it is not a sensation of pain.

There are at least two more ways in which unpleasantness and pain differ from each other, viz. the following ones:

(iv) A sensation of pain need not have any other sensory qualities besides its pain-ness, something which is explained by the fact that the "pain-ness" of the experience does not supervene on any other qualities. (This phenomenon is intimately connected to some physiological facts about us: we are equipped with certain "sense organs" whose function it is to "detect pain" (nociceptors), and there are certain (sensory) neural pathways for the transmission of "pain signals").

In the case of unpleasantness, it is different: there is no such thing as an experience that is "just unpleasant", that has no other properties (e.g., qualities) besides its unpleasantness. (And the truth of this claim is not affected by the fact that experiences can be individuated in different ways; it is simply not possible to individuate an unpleasant experience in such a way that it is "nothing but unpleasant"). One possible explanation of this circumstance is that unpleasantness is a supervenient property.

(v) The fact that every (physical) pain is a kind of mental object implies that it is possible to distinguish between the pain itself and the awareness of it, and this suggests that it is (conceptually) possible to be in pain without being aware of it. In the case of suffering, on the other hand, it is does not seem possible to distinguish the unpleasantness itself from the awareness of it. As Hare (1981) points out, "if I am suffering to a certain degree or with a certain intensity, *I must know* that I am suffering to that degree and with that intensity, and vice versa" (p 93, my italics).

Two kinds of conceptions of pleasantness (unpleasantness)

Let us now look at the most common (and most plausible) types of conceptions of pleasantness and unpleasantness, viz. (i) the hedonic-quality theories (quality-of-experience theories, or sensation models) and (ii) the relational theory (preference-theory, desire theory, motivational theory, or attitude model).

The Quality Theories

The hedonic-quality theories claim (roughly) that pleasantness and unpleasantness are sensory qualities of experiences. These so-called *hedonic qualities* are (on the theory) felt qualities, or "introspectible" qualities, i.e. all quality theories conceives of "being pleasant" as "identical with some introspectible fact of experience" (cf. Brandt (1979), p 36).

The most common form of this theory is probably the "monistic" view that Brandt attributes to Karl Duncker, viz.

> the proposal that pleasantness is a quality or attribute of a sensation, emotion, feeling, or complex of these /.../. It is said to be a quality which enables us to order the experience it qualifies in an order of more and less. It is, Duncker says, a 'tone pervading experience', incomplete by itself, essentially of something else, adjectival (ibid., p 38).

Brandt also adds that on this view, the intensity of the hedonic tone is supposed to be ranging from extreme pleasantness through indifference to extreme unpleasantness (cf. ibid., p 38). Now, as I understand it, this is not to say that a pleasant experience has more of something of which an unpleasant experience has less (what is being claimed is rather that pleasantness and unpleasantness can be measured on the same scale). That is, the view is compatible with the plausible idea that the hedonic quality which makes an experience pleasant (what Broad refers to as "the pleasant form of hedonic tone") is not the same hedonic quality which makes an experience unpleasant ("the unpleasant form of hedonic tone").

The reason why I use the term *"monistic"* to denote this type of quality theory is that the theory assumes that there is only *one kind* of "pleasantness-making" sensory quality, a kind of hedonic quality that is shared by all pleasant experiences. And the same thing is (of course) supposed to hold for unpleasant experiences: The hedonic quality which makes a certain particular experience unpleasant is (on this monistic theory) the *same* kind of quality which makes all other unpleasant experiences unpleasant.

A quality theorist need not be "monistic" in this sense, however, i.e. he does not have to assume that "hedonic tone" is a "determinable quality having two and only two determinate forms under it, viz., pleasantness and unpleasantness"; he can also admit for the possibility that there are "several different determinate forms of pleasantness and un-

pleasantness" (cf. Broad (1930), p 232). Or as Brandt (1979) puts it, a quality theorist might also claim that

> [the phrase] 'is pleasant' is a multivalent phrase having different meanings in different contexts. /.../ On this view one could claim that it is a mistake to look for some special feeling or quality of feeling, identical among /.../ [all pleasant experiences], the intensity of which correlates with our judgement of how pleasant something is (p 37).

On this *pluralistic view*, there are (instead) several different kinds of "pleasantness-making" (and "unpleasantness-making") felt qualities: our pleasures do not just differ with respect to their non-hedonic qualities, but with regard to their hedonic qualities as well. Every pleasant (unpleasant) experience is pleasant (unpleasant) in virtue of *some* hedonic quality or other, however; it is just that the hedonic quality that makes one experience pleasant need not be of the same kind as the hedonic quality that makes another experience pleasant. (This means that the quality theory is as compatible with Mill's "qualitative hedonism" as it is with Bentham's "quantitative hedonism").

To sum up, the quality theory claims that every pleasant experience is pleasant in virtue of some hedonic quality that it has; the monistic quality theory claims (on top of this) that all pleasant experiences are pleasant in virtue of having the same hedonic quality; and the pluralistic quality theory claims (instead) that the hedonic qualities which make experiences pleasant are of different kinds. Or alternatively put,

> **(Q)** An experience is pleasant if and only if (and because) it is "pervaded by" some kind of pleasant hedonic tone (quality).
>
> **(MQ1)** An experience is pleasant if and only if (and because) it is "pervaded by" *the* pleasant form of hedonic tone.
>
> **(PQ1)** Here, (Q) is supplemented by the claim that there are several different kinds (or forms) of "pleasant hedonic tone"[15].

---

[15]The claim that there are different kinds of pleasantness-making hedonic qualities should not be understood as entailing that these kinds have something in common, or that there is a genus "pleasantness-making hedonic quality" such that all elements that belong to this class have something in common. The idea is (rather) that the class of pleasantness-making hedonic qualities is a radically heterogeneous class.

The comparative counterpart to (MQ1) is

> **(MQ2)** How pleasant a certain pleasure is depends on (or more specifically, is proportional to) "how much" pleasant hedonic tone it has, or how "intense" this pleasant hedonic tone is.

So, how should we (on the pluralistic quality theory) determine how pleasant a certain pleasure is for the experiencing subject? Well, we can assume that if two pleasures are pleasant in virtue of the same hedonic tone, i.e. if they are of the same qualitative type, then the pleasure which has more of this pleasant hedonic tone is also more pleasant. But how should we compare qualitatively different pleasures with respect to pleasantness? It seems that the theory has no answer to this question. If two pleasures have no hedonic (or non-hedonic) quality in common, then there is nothing of which the one pleasure has more and the other less, and it is (for this reason) not possible (not even "in principle") to compare them with respect to their pleasantness. In short, there is (as far as I can see) no full-fledged comparative counterpart to (PQ1).

The Relational Theory

The second major type of conception of pleasantness and unpleasantness is the relational theory (desire theory, preference-theory, attitudinal model, or perhaps motivational theory). The fundamental claim that this theory makes is (roughly) that the pleasantness and unpleasantness of a person's experiences are somehow constituted by certain kinds of desires and aversions (likes and dislikes) that the person has. Or more specifically, a pleasant experience is (roughly) an experience that is, when experienced, intrinsically desired (preferred, approved of, liked, etc.) by the experiencing subject. In a similar way, an experience is unpleasant for a person if and only if he, when he has it, has an intrinsic aversion to it. In the case of pleasantness, this view can be spelled out as follows:

An experience is pleasant if and only if (and because) the following conditions are met:

> (i) The experiencing subject has some kind of pro-attitude towards the experience: he desires it, likes it, approves of it, or the like. To say that the subject desires the experience is not specific enough, how-

ever; the object of his "pleasantness-making" desire is (rather) that he has it, or that he continues to have it (cf. Brandt (1979), p 39)[16].

(ii) The experience is desired (etc.) by the experiencing subject when it is experienced.

(iii) The experience is desired in a certain way, viz. "in and for itself", i.e. intrinsically, or "as a goal", i.e. "finally"[17].

On Parfit's (1984) view, this is the conception of pleasantness and unpleasantness that the hedonist should accept (i.e. the conception which makes hedonism most plausible). He writes:

> What pains and pleasures have in common are their relations to our desires. On the use of 'pain' *which has rational and moral significance*, all pains are when experienced unwanted, and a pain is worse or greater the more it is unwanted. Similarly, all pleasures are when experienced wanted, and they are better or greater the more they are wanted. /.../ On this view, one of two experiences is more pleasant if it is preferred (p 493, my italics).

It is worth noting that the type of "rationally and morally significant pain" to which Parfit refers is more or less identical with what Hare (1981) calls "suffering". On Hare's view, "it would be self-contradictory to report suffering but claim that one did not mind it, and had no motive for ending or avoiding it, even *ceteris paribus* (p 93)". That is, every suffering is (on this view) partly constituted by a desire to put an end to it or to avoid it, or alternatively put, suffering is (at least in part)

---

[16]In connection with this, Brandt also adds another condition, viz. that it is the experience which makes the experiencing subject want its continuation. He writes: "When an experience is pleasant, the (increased) occurrent valence of [desire for] the continuation of that experience is *causally dependent* on the experience already going on /.../ [T]he occurring experience is the *differential cause*, of the increased positive valence of [desire for] the continuation of that experience" (p 40).

So, is it plausible to claim that an experience cannot be pleasant unless the subject wants it to continue? Well, it seems so, at least in the case of pleasant *states*. However, there are also pleasures that we don't necessarily want to prolong, e.g., the pleasures we get when a certain unpleasant state is removed. It can be very pleasant to have a bodily appetite satisfied (e.g., to drink when thirsty or eat when hungry), but this does not in any way imply that we (when we have it) want the experience to continue.

[17]Where intrinsic desire and final desire is not necessarily the same thing, however. Cf. the discussion in section 4.4.

a motivational state[18].

Some relational theorists have not supplemented (i) and (ii) with (iii), however, but with the following condition:

(iv) The reason why the experiencing subject desires to have the experience is (at least in part) that it has certain felt (sensory) qualities.

That is, an experience can (on this view) not be pleasant unless it is desired on certain grounds, viz. (at least in part) because it has the felt qualities it has. In a similar way, a person's experience cannot be unpleasant unless it is (in part) disliked because of its felt qualities.

An example of such a relational theorist is Sidgwick (1907). After having defined pleasure as "feeling which the sentient individual at the time of feeling it implicitly or explicitly apprehends to be desirable" (i.e. after having accepted (i) and (ii) above)[19], he then adds

---

[18]This formulation of the theory might well be sufficiently precise for our purposes, but it is important to see that it can be made even more precise, viz. if it is supplemented by a conception of desire, i.e. by a theory about what it is for a person to desire something, e.g., that a certain experience continues (this will be one of the major issues in section 4.1 and in appendix F). To see what a relational theory would look like if it would (so to speak) incorporate a conception of desire, consider Brandt's (1979) theory. On Brandt's "functional conception" of desire, desires and aversions (positive and negative "valence") are motivational states which are conceptually connected to action-tendencies (A more detailed account of Brandt's conception of desire is presented in appendix F). If we incorporate this conception into (i) above, we get: (i') The pleasantness of a pleasant experience partly consists in a "tendency not to do anything which would extinguish the experience which is pleasant; and, if the person thinks that the experience will continue only if he does something, there will be a tendency to do that. /.../ In contrast, if some kind of experience is unpleasant, the continuation of it will be negatively valenced, and hence there will be action-tendencies to do whatever is believed likely to remove, expunge, or produce its non-existence" (p 39). This explains Brandt's idea that being pleasant "is what is designated by some theoretical construct defined by its relation to behaviour" (ibid., p 36). This is how he (finally) defines the phrase "the experience *E* of the person *P* is pleasant for *P* at *t*": "'an experience of the kind *E* is going on in the person *P* at *t*; and the experience *E* is the differential cause at *t* of an increment in the positive valence of the continuation of *E* beyond *t*, or at the neonate level, of the occurrence of tendencies to act in a way likely to result in the continuation of *E*'. In short, an experience is pleasant if and only if it makes its continuation more wanted. The transposition for being unpleasant will be obvious" (ibid., pp 40-41).

[19]Cf. also ibid., p 127, where Sidgwick writes that "the only common quality that I can find in the feelings so designated [as pleasures] seems to be that relation to desire and volition expressed by the general term 'desirable' /.../. I propose therefore to define Pleasure /.../ as a feeling which, when experienced by intelligent beings, is at least implicitly apprehended as desirable or - in cases of comparison - preferable" (p 127).

- desirable, that is, *when considered merely as a feeling*, and not in respect of its objective conditions or consequences, or of any of the facts that come directly within the cognisance and judgment of others besides the sentient individual (p 131, my italics).

Another example is Broad (1930), who gives the following formulation of the relational theory:

> Is it not possible that what we have called "hedonic quality" is really a relational property and not a quality at all? Is it not possible that the statement: "This experience of mine is pleasant" just means: "I like this experience *for its non-hedonic qualities*"? /.../ On this view we should no longer divide the qualities of an experience into hedonic and non-hedonic. All its qualities would be non-hedonic. But, if its qualities were such that I liked it for them it would be pleasant, and if its qualities were such that I disliked it for them it would be painful (pp 237-238, my italics).

A third example is Nozick (1989). He writes:

> By a pleasure or pleasurable feeling I mean a feeling that is desired (partly) *because of its own felt qualities*. The feeling is not desired wholly because of what it leads to or enables you to do or because of some injunction it fulfills. If it is pleasurable, it is desired (in part at least) *because of the felt qualities it has*. I do not claim here that there is just one felt quality present whenever pleasure occurs. Being pleasurable, as I use this term, is a function of being wanted partly for its own felt qualities, whatever these qualities may be. On this view, a masochist who desires pain for its own felt quality will find pain pleasurable (p 103, my italics).

This does not mean that we need to introduce a second version of the relational theory, however, viz. for the following reason: Condition (iii) states that the experience is desired for its intrinsic properties, while condition (iv) states that it is desired for its felt qualities. Now, if we assume (plausibly) that all the intrinsic properties of an experience are (experienced) qualities, and vice versa, there is only one way in which the two conditions may differ from each other, viz. the following one: *If* there are qualities of experience which are not felt qualities, e.g., sensations of red, then there might be experiences which are pleasant accor-

ding to (iii) which are not pleasant according to (iv), e.g., looking intensely at a patch of red. So the question arises: What does "felt quality" mean in this context (and what should it mean)? Does it refer to all qualities of experience (including sensations of red), or does it only refer to those qualities which are associated with bodily feelings? Well, on my view, we should adopt the broader notion. That is, we should conceive of all (experienced) qualities as felt qualities, i.e. we should regard Nozick's and Broad's formulations of the relational theory as extensionally equivalent. The reason for this is mainly evaluative: On my view, an intrinsically desired experience can be nonderivatively good for the experiencing subject even if it lacks felt qualities in the narrow sense, and a desired experience that has such felt qualities (in the narrow sense) can be good for a person even if it is desired for its other ("non-felt") qualities. (The narrow notion of felt quality is probably more in line with ordinary language, though; it seems that in ordinary speech, we do not attribute pleasantness to an experience unless it has felt qualities in the narrow sense; cf. pp 75-79 above). In short, we should conceive of felt qualities in a way which makes conditions (iii) and (iv) extensionally equivalent.

Now, all this suggests that a pleasant experience E1 is more pleasant than another pleasant experience E2 if and only if E1 is *more desired for its (felt, non-hedonic) qualities* than E2. But what is this supposed to mean? As far as I can see, it may mean two different things, viz. (1) "E1 and E2 are both desired for their qualities, and the experiencing subject desires E1 more than he desires E2", or (2) "E1 is more desired for its qualities than E2 is desired for its qualities; the desire to have E1 is, to a larger extent, based on the fact that it has the qualities that it has". On my view, (2) seems to be the more plausible interpretation, partly because it is consistent with the fact that an experience E1 may be more pleasant (more desired for its qualities) than another experience E2, even if the subject desires E2 more than he desires E1.

To sum up, the relational theory makes the following claims:

(R1) An experience is pleasant for a person P if and only if (and because) P intrinsically wants to have it (and perhaps also that it continues) when he has it, or alternatively put, if and only if (and because) it is (in part) desired by P because of its own felt qualities.

(R2) How pleasant it is for a person to have a certain experience is a

function of (or more specifically, proportional to) how strongly it is intrinsically desired, i.e. of how much it is desired for its qualities.

This ends our survey of the most common conceptions of pleasantness and unpleasantness, or alternatively, of the most common philosophical uses of the terms "pleasant" and "unpleasant". So, let us now see what happens when these conceptions are combined with (incorporated into) pure hedonism.

Three different versions of pure hedonism

There are (as we have seen) several different conceptions of pleasantness and unpleasantness, and for each such conception, there is a corresponding version of pure hedonism. And what version of the theory a pure hedonist ends up with, this is (of course) dependent on what conception he adopts, what he thinks the terms "pleasant" and "unpleasant" refer to. So, let us now look at one conception of pleasantness and unpleasantness at the time, incorporate it into the hedonistic theory, and see what the resulting version of pure hedonism is like. We will start with the two versions of the quality theory.

The Quality Hedonisms

The Monistic Quality Theorist uses the terms "pleasant" and "unpleasant" as names of two kinds of recognizable sensory qualities, viz. the pleasant and unpleasant forms of hedonic tone. If we combine this idea with pure hedonism, we get *Monistic Quality Hedonism* (or what Parfit (1984) calls "Narrow Hedonism"). This version of the hedonistic theory can be summarized in the following two claims, of which the first is non-comparative and the second comparative:

(MQH1) An experience is nonderivatively good for a person if and only if (and because) it has a certain (pleasant) hedonic quality, i.e. if it is pervaded by the pleasant form of hedonic tone, and an experience is nonderivatively bad for a person if and only if (and because) it is pervaded by the unpleasant form of hedonic tone[20].

(MQH2) The nonderivative value that a certain experience has for

---

[20]Is this claim based on the assumption that an experience cannot be pervaded by pleasant hedonic tone and unpleasant hedonic tone at the same time? If it is, is the assumption plausible?

the experiencing subject is dependent on what hedonic qualities it has, or more specifically, it is (if we ignore the issue of duration) proportional to how pleasant or unpleasant its hedonic tone is. For example, the more pleasant hedonic tone a pleasant experience has, the better it is for the experiencing subject to have it.

On the Pluralistic Quality Theory, the term "pleasant" always refers to some hedonic quality or other, but it does not always refer to the same hedonic quality. The same thing holds for the term "unpleasant"; this term also refers to different qualities in different contexts. Now, if this theory is incorporated into pure hedonism, we end up with *Pluralistic Quality Hedonism.* This version of the hedonistic theory claims:

(PQH1) An experience is nonderivatively good for a person if and only if (and because) it has *some kind of* (pleasant) hedonic quality, i.e. if it is pervaded by some kind of pleasant hedonic tone, and an experience is nonderivatively bad for a person if and only if (and because) it is pervaded by some kind of unpleasant hedonic tone.

(PQH2) The nonderivative value that a certain experience has for the experiencing subject is (if we ignore the issue of duration) wholly dependent on what hedonic qualities it has (this is a feature that is shared by both versions of quality hedonism). But how exactly does the pluralistic quality hedonist think we should compare experiences with regard to value-for? Well, if there is no way in which two qualitatively different experiences can be compared with respect to pleasantness and unpleasantness (cf. p 86 above), then neither can they (on the assumption that pure hedonism is true) be compared with respect to value-for[21].

---

[21]This means that Pluralistic Quality Hedonism is a "methodological disaster", and it is not unlikely that these difficulties will make the pluralistic quality theorist abandon pure hedonism altogether, e.g., that it will make him reject (H5), i.e. the idea that the value of an experience for the person who has it is a function of one thing only, viz. how pleasant or unpleasant it is. Instead, he might claim (as Mill seems to have done) that the value-for of an experience is also dependent on what *kind* of hedonic tone it has. This type of modified hedonism may well be more methodologically appealing than Pluralistic Quality Hedonism: If an experience is not valuable in virtue of its pleasantness only, then incomparability with respect to pleasantness need not give rise to incomparability with respect to value-for, e.g., if some kinds of hedonic qualities are, so to speak, nonderivatively better than others.

Preference-Hedonism

If the relational theory is combined with pure hedonism, i.e. if the pure hedonist incorporates (R1) and (R2) into his theory, the resulting version of the hedonistic theory is (roughly) what Parfit (1984) calls *Preference-Hedonism*. This theory can be characterized as follows:

(PH1) An experience is nonderivatively good for a person if and only if (but not because![22]) he has an intrinsic desire to have it when he has it, or alternatively, if and only if (but not because) he wants to have the experience (when he has it) because of its felt qualities. In a similar way, an experience is nonderivatively bad for a person if and only if he has an intrinsic aversion to having it when he has it, i.e. iff (but not because) he wants to get rid of the experience because of its felt qualities.

(PH2) How nonderivatively good it is for a person to have a certain pleasant experience is (if we ignore the issue of duration) proportional to how strongly the person (intrinsically) desires to have it when he has it, or alternatively, to how much it is desired for its own felt qualities. That is, the stronger an experience is desired, and the higher the extent to which this desire is based on the fact that the experience has the felt qualities it has, the better it is for the experiencing subject to have the experience. (In a similar way, the degree to which a certain unpleasant experience is bad is a function of how strongly it is "negatively valenced" by the experiencing subject). This means that it is nonderivatively better for a person to have a certain pleasant experience E1 than to have another pleasant experience E2 if and only if E1 is more desired for its felt qualities than what E2 is desired for *its* felt qualities.

This theory should be carefully distinguished from the version of the desire theory that I call *the experience-oriented Success theory* (a theory which will be further discussed in section 5.2.1). On the object interpretation of this theory, the only type of situation that is nonderivatively good for a person is that he has an experience that he wants to have, and the only thing that is nonderivatively bad for a person is that

---

[22]The reason why a certain experience is good for a certain person is still that it is pleasant, and not that it is desired! Cf. the discussion on the experience-oriented Success Theory below.

he has an experience that he wants not to have. The experience-oriented Success theory also claims that it is nonderivatively better for a person to have an experience E1 than it is to have another experience E2 if and only if he intrinsically prefers E1 to E2.

Now, the object interpretation of this theory might appear identical with Preference-Hedonism, but this is not really the case. It is true that both theories accept (H1), i.e. "the experience requirement", but there are also differences between the theories.

First, both theories imply that a certain experience cannot be nonderivatively good for a certain person unless he has an intrinsic desire to have it, but whereas the experience-oriented Success Theorist claims that it is nonderivatively good for a certain person to have a certain experience if and only if he has an intrinsic desire to have the experience, the Preference-Hedonist claims that it is nonderivatively good for a certain person to have a certain experience if and only if he has an intrinsic desire to have the experience *when he has it* (and so on)[23].

However, it can be argued (and it will, viz. in section 4.3) that all

---

[23]Even though this feature is included in Parfit's (1984) own "definition" of Preference-Hedonism (cf. the quotation on p 87 above), he sometimes tends to ignore it totally. Look at what he writes: "Near the end of his life Freud refused pain-killing drugs, preferring to think in torment than to be confusedly euphoric. Of these two mental states, euphoria is [in the ordinary sense of 'pleasure'] more pleasant. But on Preference-Hedonism thinking in torment was, for Freud, a better mental state. It is clearer here not to stretch the meaning of the word 'pleasant'. A Preference-Hedonist should merely claim that, since Freud preferred to think clearly though in torment, his life went better if it went as he preferred" (pp 493-494).

Here, Parfit is most certainly wrong about what a Preference-Hedonist would say about Freud's case. Let us (for the sake of argument) make the rather dubious assumption that thinking in torment is an experience, and that Freud's preference for thinking in torment to being confusedly euphoric is a preference for one type of experience to another one. Now, it is very likely that Freud had this preference at *one* particular time, supposedly when he was thinking in torment, or when he was neither thinking in torment nor being confusedly euphoric, and *not* when he was confusedly euphoric. In short, the two Freudian desires whose respective strengths are being compared are most probably not the desire to think in torment when thinking in torment and the desire to be confusedly euphoric when being confusedly euphoric.

Moreover, if we add the requirement (which Parfit sometimes seems to forget about) that the relevant desires are intrinsic, it is rather clear that a Preference-Hedonist would regard it as better for Freud to be confusedly euphoric. In fact, he would even allow for the possibility that it was nonderivatively bad for Freud to think in torment. It is, after all, not likely that Freud desired to think in torment *because of what it felt like*. (However, an enjoyment-hedonist (cf. note 10) would probably regard thinking in torment as better for Freud).

plausible desire theories regard both prospective and retrospective desires as irrelevant. Now, if the experience-oriented Success Theory we have in mind shares this feature (if it gives weight to now-for-now desires only), we can see that the object interpretation of this theory is extensionally equivalent with Preference-Hedonism: the two theories attribute nonderivative value-for-P to exactly the same facts. This does not make the two theories identical, however. The experience-oriented Success Theory does not (unlike the hedonistic theories) claim that it is good for a person to feel pleasure, and it does not make any claims whatsoever about what is pleasant and unpleasant. What it claims is that it is nonderivatively good for a person P to have an experience E if and only if (and because!) P has an intrinsic now-for-now desire to have E (and so on), i.e. that every good experience is good in virtue of being intrinsically desired by the experiencing subject. The Preference-Hedonist, on the other hand, claims that good experiences are good because they are pleasant, and *not* because they are intrinsically desired. (As I see it, this is not changed by the fact that he also makes another claim, viz. that all pleasant experiences are pleasant in virtue of being intrinsically desired).

This concludes the survey of the different versions of pure hedonism.

A critical discussion of the different conceptions

Let us now try to determine which theory of pleasantness is (in this context) most plausible, or alternatively put, which version of pure hedonism is most plausible. Now, the only way in which this can be done is by examining a number of arguments that has been (and can be) given for and against the respective theories. These arguments are (as we will see) of different kinds: Some of them appeal to our evaluative intuitions, while others appeal to our semantical, ontological, or phenomenological intuitions.

It should be pointed out that the only arguments which are (strictly speaking) arguments for and against the different versions of pure hedonism are the "evaluative" arguments; all the other arguments are really arguments for and against the different conceptions of pleasantness on which the different hedonisms are based.

The background assumption on which the arguments are based can be formulated as follows: A conception of pleasantness and unpleasant-

ness cannot be plausible unless it meets the following three requirements, viz. (i) it should use the terms "pleasant" and "unpleasant" in a way that has "rational and moral significance", or more specifically, it should "make" the hedonistic theory as plausible as possible; (ii) it should be ontologically sound, i.e. it should not presuppose the existence of something which does not exist (the terms "pleasant" and "unpleasant" should refer to something that exist); and (iii) the terms "pleasant" and "unpleasant" should be used in a way that is "in line with" ordinary speech, that does not deviate too much from ordinary speech[24].

Let us now look at the arguments for and against the different conceptions. There will first be a more general discussion of the pros and cons of the quality theory, a discussion which mainly consists of semantical and/or phenomenological arguments. This discussion is also (to a large extent) a discussion of the pros and cons of the relational theory: In the present context, every argument for the quality theory is, after all, an argument against the relational theory (and vice versa).

The Quality Theory vs. the Relational Theory: A general discussion

The arguments in this section appeal to our semantical and/or phenomenological intuitions. Most of them are "ordinary language" arguments, and are (as such) based on requirement (iii) above.

Let us first look at some arguments that can be given for the quality theory and against the relational theory:

(1) We often talk about pleasantness as if it is something we can feel, e.g., as when we say that we "feel pleasure", or that we feel how pleasant a certain on-going experience is. This shows that the quality theory is (somehow) embedded in ordinary speech, and in the common sense conception of pleasantness.

(2) We tend to believe (with Hare (1981)) that "if I am suffering to a certain degree or with a certain intensity, I must know that I am suffering to that degree and with that intensity, and vice versa" (p 93); this

[24]A conception of pleasantness and unpleasantness need not (at least not in this evaluative context) be "methodologically appealing" in order to be plausible, however. Suppose that there are two alternative conceptions of pleasantness, C1 and C2, and that C1 "makes" pleasantness measurable in a stronger sense than C2 does. This is no reason whatsoever for regarding C1 as more plausible than C2.

is how we use the word "suffering". So, if this is so, how do I know that a certain experience of mine is pleasant (or unpleasant) to a certain degree? The quality theory gives a simple and plausible explanation of this alleged fact, viz. it tells me that the reason why I know that a certain experience is pleasant to a certain degree is that I *feel* that it is pleasant to that degree. For the relational theory, it is not as easy to come up with such a good explanation, e.g., it would have to assume that we can have introspective knowledge of our "experiential desires".

(3) We tend to think that if a certain person intrinsically wants a certain on-going experience to continue, then the reason for this is probably that the experience is pleasant. Or alternatively put, we think it makes sense to say about a person that he intrinsically likes to have a certain experience *because* it is pleasant, and we also tend to find such statements informative. This "observation" supports, if correct, the quality theory. It is incompatible with the relational theory, however, and for the following reason: The claim that every pleasant experience is pleasant in virtue of being intrinsically liked (etc.) by the experiencing subject implies that it is tautological (and not at all informative) to say that someone intrinsically likes a certain on-going experience because it is pleasant.

The next argument is against the quality theory and for the relational theory:

(4) It is hard to deny that there is a very intimate connection between two kinds of facts, viz. facts about pleasure and facts about certain intrinsic desires. For example, it seems that whenever a person feels pleasure, he also has an intrinsic desire to feel what he is feeling[25] (but not necessarily vice versa; cf. below). So the questions arise: Why is it that we always (or almost always) like the experiences we find pleasant? And what is the nature of the connection between the fact that a certain person has a certain pleasant experience, on the one hand, and the fact that this person, when he has the experience, intrinsically likes to have it, on the other?

Consider the case of sensation: Suppose that I find the taste of cognac pleasant, and that I intrinsically like the sensation when it is going on. How is the fact that the sensation is pleasant connected to the fact that

---

[25]Note that this claim is compatible with the fact that some people have an "non-intrinsic" aversion against pleasure, e.g., that someone dislikes a certain pleasure on moral grounds, or for instrumental reasons.

I like the sensation? On the relational theory, the pleasantness of the sensation is constituted by my liking to have it when I have it, i.e. the relation is conceived of as conceptual (or logical). On the quality theory, on the other hand, the relation is (rather) contingent, e.g., causal (cf. internalist vs. externalist views on the nature of the connection between evaluation and motivation).

This means that the relational theorist is in a better position to explain why we always (or almost always) like the sensations that we find pleasant. It is more difficult for the quality theorist to explain why we so often happen to like the pleasant sensations we have, but far from impossible. So, why do we (according to the quality theorist) intrinsically prefer pleasant sensations to unpleasant sensations, and why do we want pleasant sensations rather than sensations of blue? Well, all he can do here is to claim that "that's how we *happen* to be constituted". But is it really a contingent fact about us that we want to avoid suffering?[26]

Here are two similar arguments, but now directed against the relational theory:

(5) It seems plausible to assume that whenever a person has a pleasant experience, he also has an intrinsic desire to have the experience, and that this fact is of a conceptual nature. But the relational theorist makes a stronger claim, i.e. he also claims that whenever a person has an intrinsic desire to continue having a certain on-going experience, this experience is a pleasant experience. This is far from obvious, however. Or more precisely, the stronger claim may well be incompatible with our common sense conception of pleasantness.

(6) On the relational theory, the pleasantness of every pleasant experience is constituted by the fact that the experiencing subject intrinsically desires to have the experience when he has it. Now, consider the case of emotion: Here, the relational theory claims that the *only* reason why a pleasant emotion is pleasant is because it is liked (etc.) when it is going on. However, it seems that an emotion may also be pleasant for other reasons, e.g., there are cases (think of the type of happiness described in note 13) where the pleasantness of an emotion is (at least in part) constituted by the pleasantness of its component sensations[27]. And if

---

[26]As Hare claims, "it would be self-contradictory to report suffering but claim that one did not mind it, and had no motive for ending it or avoiding it" (cf. p 87 above).

[27]Where the relational theorist may (of course) be right about what the

this is so (if it is sometimes the case that a pleasant emotion is pleasant in virtue of having a pleasant sensory content), the relational theory must be wrong.

The next argument, which is directed against the quality theory, also has to do with the pleasantness and unpleasantness of more complex experiences, e.g., emotions or moods.

(7) So, what would the quality theorist say about the pleasantness and unpleasantness of emotions and moods? Well, he would have to claim that an emotion (or mood) is pleasant or unpleasant in virtue of its sensory content (felt qualities) *only*, e.g., that the only reason why a certain pleasant emotion is pleasant is because it contains certain pleasant sensations. Let us call this idea *the sensory component thesis*.

We have assumed (cf. argument (6) above) that there is some truth in this thesis, i.e. it seems that pleasant emotions may sometimes be pleasant in virtue of having a pleasant sensory content. However, there are (it seems) cases of pleasant emotion where the pleasantness of the emotion is also constituted by other things, and not just by the pleasantness of its component sensations. To see this, consider an emotion that can (depending on the context) be both pleasant and unpleasant, e.g., a negative emotion like anger or grief. Let us now ask what the difference between pleasant anger and unpleasant anger consists in. Is it really *nothing but* the sensory components of the two emotions that make them differ in this way? I think not. It is true that pleasant anger does not feel exactly the same as unpleasant anger, but the "sensory similarities" between two angers are (nevertheless) greater than the "sensory differences". This means (I think) that the difference in feeling between the two angers does not suffice to explain why one of them is pleasant while the other is unpleasant. Instead, we can assume that the reason why the first anger is pleasant is (in part) that the angry subject likes to have it when he has it, and the reason why the second anger is unpleasant is (in part) that the angry subject dislikes to have it. And if this is correct, the quality theory is wrong.

So, what conclusions can we draw from this semantical and/or phenomenological "discussion"? Well, it seems that our semantical intuitions are somewhat "ambiguous"; most of them support the quality theory, but there are also a few that support the relational theory. The

---

pleasantness of these constituent sensations consists in.

most plausible explanation of this circumstance is probably that the terms "pleasant" and "unpleasant" have several different meanings in ordinary speech, or more specifically, that the quality theory and the relational theory are both "embedded" in ordinary speech. This suggests that both types of theories meet requirement (iii), i.e. that neither of them (especially not the quality theory) deviate too much from the common sense conception embedded in ordinary speech. If all this is correct, we cannot really decide between the two theories on semantical grounds. But even if one of the theories were superior in this respect (if it were much more in line with ordinary speech), we should not put too much emphasis on this fact; ordinary language is (after all) defective in several ways, i.e. it may contain (or be based on) false ontological assumptions, and so on.

So, let us now look at a different set of arguments, arguments which do not appeal as much to our semantical intuitions. The first of these arguments is directed against the Monistic Quality Theory, and the second against the Pluralistic Quality Theory.

A heavy objection to the Monistic Quality Theory

The major objection to the Monistic Theory is this: The theory assumes that there is a kind of hedonic sensory quality which is shared by all pleasant experiences (and which makes them pleasant), and that there is another kind of hedonic quality which makes all unpleasant experiences unpleasant. This ontological (and phenomenological) assumption is false, however; the two hedonic qualities postulated by the theory do not exist[28]. The argument against the assumption is simple: The idea is that if we consider "the variety of pleasant experience", i.e. if we consider how much our pleasures differ from each other, then we will realize that they do not have any felt quality (hedonic or non-hedonic) in common. Parfit's (1984) criticism of what he calls "narrow hedonism" (the version of the hedonistic theory which is based on the monistic quality theory) is a good example of this type of argument:

Narrow hedonists assume, falsely, that pleasure and pain are two dis-

[28]It is important to note that this is a phenomenological claim, and that it should (as such) be carefully distinguished from the methodological idea that it is not possible to measure the pleasantness or unpleasantness of all experiences on the same pleasantness-unpleasantness scale.

tinctive kinds of experience. Compare the pleasures of satisfying an intense thirst or lust, listening to music, solving an intellectual problem, reading a tragedy, and knowing that one's child is happy. These various experiences do not contain any distinctive common quality (p 493).

Griffin's (1986) criticism of Monistic Quality Hedonism is of a similar kind (but it also contains an evaluative element: it also appeals to our intuitions about what has "utility"). On the Monistic Quality Hedonist's "psychological account of 'utility'", Griffin writes,

> [p]leasure or happiness is presented as a 'state of feeling', and pain or unhappiness as a feeling on the same scale as, and the opposite of, pleasure or happiness. And the utilities of all our experiences are supposed to be determinable by measuring the amount of this homogenous mental state that they contain.
>
> The trouble with thinking of utility as *one* kind of mental state is that we cannot find any one state in all that we regard as having utility - eating, reading, working, creating, helping. What one mental state runs through them all in virtue of which we rank them as we do? /.../
>
> So, if the mental state account [i.e. hedonism] takes this simple form, the objections to it are insurmountable (p 8).

How can the Monistic Quality Theorist meet this objection? Well, he can always refuse to accept the objection, and cling to his theory on "intuitive grounds". He can admit that the "pleasantness-making" sensory quality may seem elusive, but still claim (like Broad once did) that hedonic tone is a quality "which we cannot define but are perfectly acquainted with" (cf. Brandt (1979), p 38). And how we should respond to this, I don't know[29].

On my view, the objection given to the Monistic Quality Theory is a heavy objection, and the theory should therefore be rejected. So, let us now see whether the Pluralistic Quality Theory is more plausible.

---

[29]Cf. the following case: How should we respond to someone who, after having read the relevant paragraphs in Wittgenstein's *Philosophical Investigations*, exclaims: "Yes, but all games must be games in virtue of something, and even though we may not be able to specify what this is, we are perfectly acquainted with it. We do, after all, know how to use the word 'game'!".

Two objections against the Pluralistic Quality Theory

As far as I can see, there is no knock-down argument against the Pluralistic Quality Theory; it does not get hit by the objection against the monistic theory, it is hard to prove that the hedonic qualities it postulates do not exist, and it gets some support from ordinary speech (especially from (1) and (2) above). It does give rise to certain difficulties, however.

The most difficult problem that a Pluralistic Quality Theorist has to face is (of course) this: The theory assumes that the class of pleasant experiences is a ("qualitatively speaking") radically heterogeneous class, i.e. that there are different forms of pleasantness, forms which have (qualitatively speaking) nothing in common. This assumption gives rise to the following problem: What is it that makes all these different kinds of pleasant experiences *pleasant*? In virtue of what are all these "forms of pleasantness" forms of *pleasantness*? Now, it is doubtful whether the Pluralistic Quality Theorist can solve this problem, i.e. it is likely that he must (as Brandt (1979) suggests), in principle, "leave unanswered the questions why we happen to apply 'is pleasant' to this particular set of experiences" (p 37).

Another difficulty has already been pointed out, viz. on p 86: The fact that there are (on the theory) several different kinds of pleasant (unpleasant) hedonic tone makes it hard (or even impossible) to tell how our experiences should ("in principle") be compared with respect to pleasantness and unpleasantness. This means that if a pure hedonist accepts this conception of pleasantness, then it will be immensely difficult for him to formulate his comparative claims (his claims about betterness-for) in a coherent way. (It is easier for the hedonist to handle this difficulty if he switches to a modified version of the theory, however; cf. note 21 above)[30].

[30]That is, the idea is not just that pleasantness cannot be measured (given a certain criterion on how it should be measured), but that it may even be impossible to formulate such a criterion in a coherent way. This does not necessarily count against the pluralistic quality theory, however.

Why we should prefer the Relational Theory to the Pluralistic Quality Theory

The obvious alternative to the Pluralistic Quality Theory is the Relational Theory, so let us try to figure out which type of theory that is most reasonable.

As I see it, there is only one reason for preferring the Pluralistic Quality Theory, viz. that it (*qua* quality theory) tends to get more support from ordinary speech (especially from (1) and (2) above). However, there are several reasons for regarding the Relational Theory as the more plausible theory:

(i) It is as ontologically sound as it can be, i.e. it does (most definitely) not presuppose the existence of something which does not exist. In this respect, it is superior to the Pluralistic Quality Theory. It is (as we have seen) hard to prove that the hedonic qualities postulated by the latter theory do not exist, but their existence can most certainly be doubted.

(ii) The Relational Theory does not give rise to any of the difficulties that the Pluralistic Quality Theorist has to face. First, the Relational Theory has a simple answer to first of the Pluralist's problems, viz. "What is it that makes all these qualitatively different kinds of pleasant experiences pleasant?". His answer is (of course): "They all stand in the same relation to the desires (or likes) of the experiencing subject". As Griffin (1986) writes:

> Suppose we said that utility [pleasure] consisted of several different mental states. What then would make them into a set [e.g., make them comparable with respect to their pleasantness]? The obvious candidate would be desire; we could say /.../ that utility combines a psychological [quality] element and a preference element. 'Utility', we could say, is 'desirable consciousness', meaning by 'desirable' either consciousness that we actually desire or consciousness that we would desire if we knew what it would be like to have it (pp 9-10).

The Relational Theorist avoids the second difficulty as well. He has (as we know) a simple and coherent view on how our experiences should (in principle) be compared with respect to pleasantness and unpleasantness, and it will therefore be easy for the Preference-Hedonist to give a clear and coherent formulation of his comparative claims.

(iii) The last reason why we should prefer the Relational Theory to the Pluralist Quality Theory is evaluative, viz. because the Relational

Theorist uses the terms "pleasant" and "unpleasant" in a way that has more "rational and moral significance". Or in other terms, Preference-Hedonism is a more plausible version of the Hedonistic Theory than Pluralistic Quality Hedonism. To argue for this view, we have to construct an example where the two theories disagree. This is kind of difficult, however. First, we have to assume that there are such things as hedonic qualities, and we have to imagine two experiences, E1 and E2, that have the same kind of hedonic quality. We then have to assume that E1 has more of this hedonic quality than E2, but that the experiencing subject intrinsically prefers E2 (when he has it) to E1 (when he has it). (It is hard to be more concrete than this). We then have to ask which experience is better for the subject to have, E1 or E2. On my tentative view, E2 is the more valuable experience, and preference-hedonism is (for this reason) the more plausible view[31].

This ends the chapter on how pure hedonism should be formulated (but see also appendix E, where there is a discussion of the issue of duration). Let us now try to find out whether hedonism is a plausible theory of prudential value.

[31]At this point, we may also add the following "argument" against quality-hedonism (monistic *or* pluralistic), viz. that quality-hedonism is an objectivist theory, and as such, it suffers from a number of defects (cf. chapter 7).

Chapter Three

Is hedonism a plausible theory of prudential value?

A critical discussion of the hedonistic theory

The main purpose of this chapter is to find out whether hedonism is (in any of its versions) a plausible theory, where the term "plausible" may either mean "valid" (e.g., "true") or "justified" (or "well-founded"). The central question here is (of course) whether any version of hedonism is a valid theory. This question might not be possible to answer, however. First, there might not be such a thing as valid theory of prudential value (where validity can be clearly distinguished from "well-foundedness"), and second, even if there is such a thing, it might not be possible to determine whether the hedonistic theory is valid or not. In any case, it is clearly not possible to gain knowledge about its validity in any direct way (it is, after all, a general theory). So, what we will focus on here is (rather) the question of whether hedonism (in any of its versions) is a justified (or well-founded) theory, i.e. whether there are *good reasons* for its truth and/or for accepting it.

More specifically, there are two questions we want to answer: First, is any version of hedonism justified, i.e. can (H1) and (H2) be justified? And second, if the answer is yes (if we can assume that some version of hedonism is justified); which version of the theory is most well-founded, pure hedonism or some version of modified hedonism? Or in other terms, are (H3) and/or (H4)-(H6) well-founded claims?

To be able to answer these questions, we have to look at a number of arguments that has (or can) be given for and against the theory, and then ask ourselves whether these arguments are good arguments. For example, are there good reasons for regarding (H1)-(H6) as true? Are there good reasons for regarding any of these claims as false, i.e. can pure hedonism be refuted?

What we are primarily interested in here is (of course) objective justification (cf. section 1.3 above), i.e. whether there are good reasons

(period) for or against the theory. This might not be possible, however. First, it might be impossible "in principle" to determine whether a theory of prudential value is justified (refuted) in the objective sense. As Bergström (1990) suggests, it seems that a statement cannot be objectively justified or refuted unless it has truth-value, and that value statements have truth-values is something that can be doubted. Second, even if there is such a thing as "being justified (refuted) in the objective sense" in this area, it is probably "practically impossible" to determine whether a certain theory of prudential value (e.g., pure hedonism) is objectively justified or not. Philosophers simply have too different opinions about what counts as a good reason for and against a certain theory, and it seems almost impossible to construct arguments that everyone can accept. This suggests that it *might* be better if we focus on the subjective side of the issue, i.e. if we (instead) ask whether it is rational for us (me, you, etc.) to believe that hedonism is true. Are we (I, you, etc.) justified in regarding the hedonistic theory (in any of its versions) as valid (or invalid)?

Corresponding to the distinction between objective and subjective justification, there are two ways in which we can conceive of the arguments that are given for and against a certain theory (e.g., hedonism): We can regard them "objectively", as reasons for or against the truth of the theory, but we can also regard them "subjectively", as attempts to convince a certain audience (an audience that is already equipped with certain "intuitions") that the theory is true or false[1]. Insofar as it is

[1]This is not to deny that most arguments for and against hedonism are (in fact) *attempts* at objective justification or refutation. For example, it seems quite clear that what Nozick was actually trying to do when he offered his famous experience machine argument was to refute the hedonistic theory (objectively). That is, his argument was not just an attempt to make us believe (with rational means, by appealing to our intuitions) that hedonism is false. But what he was actually trying to do is one thing, and what he was actually doing (and succeeded in doing) is another. If there is no such thing as being an objectively justified (or refuted) theory of prudential value, then it is better if we conceive of his argument as a "subjective reason" rather than as an "objective reason". Moreover, there are also a number of arguments which seem to aim at convincing rather than "proving" (or establishing truth), arguments which purport to make hedonism more or less plausible for a certain audience (equipped with certain intuitions). For example, hedonists normally try to show that it is *not* nonderivatively good for a person to be autonomous, to have friends, to have his desires fulfilled, to accomplish something, or the like (i.e. that (H2) is well-founded), but few hedonists try to show that it is nonderivatively good for us to feel pleasure (that (H3) is justified). The reason for this is most probably that the hedonist's primary

possible, we will conceive of the arguments as "objective reasons", and evaluate them accordingly. But at times, it might be more appropriate to regard them as "subjective reasons". And when we do this, we will have to evaluate them in a somewhat different way, e.g., we will have to ask whether they actually succeed in convincing a certain audience in a rational way.

Now that we are aware of these things, let us look at a number of arguments that can be given for and against hedonism. Let us start with the pro-arguments.

3.1. Arguments for hedonism

Before we look at the pro-arguments themselves, it is important to make a few general remarks about what is needed in order to justify a theory of prudential value in general, and hedonism in particular. What would a good argumentation for hedonism have to be like?

Arguing for (justifying) hedonism: some general remarks

What the hedonist needs to establish is (roughly) that pleasantness and unpleasantness have a very special evaluative status. More specifically, the key claims that he has to argue for are (H2) and (H3). What he needs to show is that there are good reasons for regarding these two claims as true.

First, he needs to show that (H3) is well-founded, i.e. that there are good reasons for believing that it is always nonderivatively good for us to have pleasant experiences, and that it is always bad for us to suffer. It is normally not necessary to give arguments for (H3) in order to convince an anti-hedonist, however (see note 1)[2].

Second, he needs to show that (H2) is justified, i.e. that there is nothing else besides feeling pleasure or displeasure that is nonderivatively good or bad for a person. This means that every argument for

---

goal is to convince the anti-hedonist. The ordinary anti-hedonists already tend to accept the view that it is nonderivatively good for us to feel pleasure, and this means that the hedonist can put his efforts elsewhere. But if his only aim were to justify his theory objectively, he would have to argue for (H3) as well.

[2]With a few exceptions, that is. For example, it might be necessary to convince some anti-hedonists that it is nonderivatively good for a person to feel sadistic joy.

(H2) will also be an argument against some alternative theory of prudential value. That is, in order to justify hedonism, it is not sufficient to show that the positive claim (H2) is justified; the hedonist also has to refute all the alternative theories of prudential value. In practice, this is where the great challenge lies. It is as hard for a hedonist to make an non-hedonist believe (by rational means) that hedonism is true as it is for a non-hedonist to convince an hedonist that hedonism is false.

So, what type (or types) of arguments would count as good reasons for the truth of (H2) and (H3)? If we assume that there are in fact good reasons for these claims, what would these reasons have to be like?

The three most common types of pro-arguments

(1) Arguments for (H2), and against some alternative theory, are often of the following kind: The hedonist takes some non-hedonistic belief (e.g., the belief that it is nonderivatively good for us to engage in creative activity) and tries to explain its occurrence in "hedonistic terms". The reason why some people tend to regard it as nonderivatively good for us to be "creatively active" is (on this type of view) that creatures like us tend to take pleasure in it; if it would not give us pleasure, we would probably not regard it as good at all[3].

Arguments of this type are far from conclusive, but this does not mean that they are bad. In any case, it is clear that such arguments might be highly effective, or more specifically, if a person would come to believe that all his non-hedonistic beliefs can be explained in this way, then it is likely that he would convert to hedonism.

---

[3]Cf. Smart's (1973) arguments for pure hedonism and against modified hedonism: (i) The reason why some of us (e.g., Mill) believe that complex and intellectual pleasures are nonderivatively better than sensual and simple pleasures is that the former are often fecund while the latter do not only lack fecundity, but are actually (often) the reverse of fecund (cf. p 17-18). (ii) The reason why some of us (e.g., Moore) tend to think that sadistic pleasure has "no intrinsic value at all, or perhaps even a *negative* intrinsic value" (ibid., p 25) is that we feel a distaste for the consequences of sadism, and not an immediate distaste for sadism as such: "Our repugnance to the sadist arises, naturally enough, because in our universe sadists invariably do harm" (ibid., p 25). In both cases, the modified hedonist is (somehow) characterized as confused, conceptually or psychologically: The reason why he is not a pure hedonist is either that he blurs the distinction between final and instrumental values, or that he is bad at introspection.

(2) Arguments of type (1) are often supplemented by arguments of a similar type. These arguments can be characterized as follows[4]: The hedonist considers an opposing view, such as the idea that it is non-derivatively good for people to be engaged in creative activity. He then asks (rhetorically): "Would it be good for a person to engage in creative activity, if it brought no enjoyment at all to him to do so, and if he did not have the slightest desire that the experience of doing so continue?"[5]. He gives a "no" answer to this question, and then concludes that the value of engaging in creative activity must lie in the pleasure it produces, and in it arousing a desire that the experience of it continue.

As Parfit (1984) points out, "[t]his reasoning assumes that the value of a whole is just the sum of the values of its parts" (p 501), i.e. if this atomistic assumption is not made, the alleged fact that a certain whole (e.g., P-is-happy-because-he-wants-to-engage-in-creative-activity-and-now-he-does-it) can not have final value for P unless it contains pleasure as a part does not give any support to the hedonistic idea that the value-for of the whole *resides* in this part. This assumption is not plausible, however, and we therefore have to conclude that (b) is (at least *qua* argument for pure hedonism) a bad argument.

The other premise of the argument is not implausible, however, i.e. it may well be the case that a situation cannot have final value for a person unless it includes pleasure as a component. If we reject atomism but accept this premise, we will end up with the following type of view:

> We might then claim that what is best for people is a composite. It is not just their being in the conscious states that they want to be in. Nor is it just their having knowledge, engaging in rational activity, being aware of true beauty, and the like. What is good for someone is neither just what Hedonist's claim, nor just what is claimed by Objective List Theorists. We might believe that if we had *either* of these, *without the other*, what we had would have little or no value (Parfit (1984), p 502).

Even though this argument is not an argument against pure hedonism

---

[4]The description that follows is a reconstruction of what Parfit (1984) writes (somewhat confusedly) on p 501.

[5]At this point, he might also invite us to make comparisons of the following type: "Suppose that there is this person who suffers from being creative, but takes pleasure in performing routine tasks. Is it really nonderivatively better for this person to be creative?".

itself, but only an argument against an argument for pure hedonism, it does weaken the hedonist by "disarming" him. If we take this weapon away from him, what has he got left? How can he support his thesis now?

(3) Well, he can also argue as follows: "Let us (again) consider the idea that it is nonderivatively good for people to be engaged in creative activity. This view implies that a person can become better off even if there is an increase in his suffering, viz. if there is a large enough increase in creative activity. This is implausible, however, and the view should therefore be rejected. Moreover, it seems that whatever alleged value we replace creative activity with, we get the same result: There simply is no thing such that a decrease in a person's level of happiness can be compensated by an increase in this thing".

This argument might strike some people as convincing, but it is not as strong as it may seem. First, it is not really an argument for pure hedonism. It is true that the idea that every decrease in happiness is also a decrease in well-being implies that the pure versions of the opposing views must be wrong, but the idea is not inconsistent with all the opposing views, e.g., it is perfectly consistent with Parfit's idea that "what is best for people is a composite". And second, it is doubtful whether the argument can convince anyone who isn't already a hedonist. Consider this person who has a strong intrinsic desire to learn more about himself. Now, suppose he learns something about himself which makes him suffer a little more. This does not necessarily make him worse off, especially not if he himself intrinsically prefers painful knowing to painless ignorance.

On subject-oriented vs. object-oriented justification of value-for-statements, and why subject-oriented justification is necessary

Even though some of the arguments above might be sufficient to convince some non-hedonists, they can never constitute a full objective justification of the hedonistic theory. First, a complete (and satisfactory) justification of (H2) must be (in part) "subject-oriented"[6]. Second, the

[6]Cf. appendix B. It is also worth reminding ourselves that the distinction between subject-oriented and object-oriented justification has nothing to do with the

hedonist also has to show that there are good reasons for believing that it is always nonderivatively good for us to feel pleasure (etc.), and this justification must also be (to some extent) "subject-oriented". In short, it is not just that both (H2) and (H3) need to be justified; they also need to be justified in a "subject-oriented" way, the hedonist also has to tell us what it is about *us* (e.g., our human nature) that makes hedonism plausible. Let us elaborate further on this idea.

What the hedonist is trying to justify (in this context) is a number of general value-for-statements, however, e.g., he tries to show that there are good reasons for believing that nothing but pleasant experience is nonderivatively *good for us*. Now, the fact that value-for is a relation between an object and a subject strongly suggests that if a certain object is nonderivatively good for a certain subject, then this is so because there is a "fit" between the object and the subject: The object has certain properties, the subject has certain other properties, and the two sets of properties match each other well (cf. appendix B). In the present case, this means that general value-for-statements of the form "all facts of type X are nonderivatively good for all human beings" can not be justified by referring solely to what X's are like; one must also refer to what human beings are like, e.g., to "human nature". (In the case of value-period, it is different. An attribution of final value to an object is justified by referring solely to intrinsic (or perhaps relational) features of this object, i.e. by referring solely to "the nature" of this object).

Let us now make the following distinction between object-oriented and subject-oriented justification: A particular statement of the form "the fact X is good for the person P" is *justified in an object-oriented way* if the justification refers to features of the object (X), i.e. if it takes the form "X is good for P *because* X is constituted in such-and-such a way". A *subject-oriented justification* of a particular value-for-statement, on the other hand, "appeals" to features of the subject (P), i.e. it takes the form "X is good for P *because* P is constituted in such-and-such a way". The point made above can now be formulated in the following way: A goodness-for-statement can not be completely justified unless the justification is both object-oriented and subject-oriented, i.e. unless it takes the form: "X is good for P *because* X is constituted in such-and-such a

---

distinction between objective and subjective justification. (It is important that these two distinctions are kept separate).

way *and because* P is constituted in such-and-such a way".

This does not necessarily apply to general statements of the form "all facts of type X are nonderivatively good for all human beings", however, viz. for the following reason: When we see such statements, we already know what it is about X's that is supposed to make them good for us, viz. that they are X's, that they fall under the general description "being an X (e.g., being pleasant)". This means that justifications of general value-for-statements might not have to be object-oriented at all, not even in part. Instead, the central requirement is this: *A general goodness-for-statement can not be completely justified unless the justification is subject-oriented*, i.e. unless it takes the form "All facts of type X are nonderivatively good for all human beings *because* human beings are constituted in such-and-such a way" (assuming that X does not contain any essential reference to our intrinsic features; cf. appendix B, including notes 38 and 39).

This implies that the hedonistic theory can not be completely justified unless it can be established what it is about us (our nature, our constitution) that makes it nonderivatively good for us to have pleasant experiences, and so on. So, what kinds of subject-oriented justifications can be given here? What kinds of human features might be of relevance in this (hedonistic) context?

Well, it hardly makes sense to say that it is good for us to feel pleasure because we are sentient beings (because we have a capacity for pleasure and pain), or because we have a need for pleasure, or because feeling pleasure constitutes a realization of "the human potential". As I see it, a subject-oriented justification of hedonism can only make sense if it appeals to our intrinsic desires or evaluations; either to our *actual* (intrinsic) desires or evaluations, or to our *hypothetical* (intrinsic) desires or evaluations (i.e. to what we would intrinsically want or value under certain circumstances). On this view, there are four major subject-oriented ways in which hedonism might be justified, viz. (1) by appealing to our actual desires; (2) by appealing to our actual evaluations; (3) by appealing to our informed or rational desires; or (4) by appealing to our informed or rational evaluations. Let us take a closer look at these different kinds of subject-oriented justifications.

Subject-oriented arguments for hedonism

(1) The first type of subject-oriented argument for hedonism is "the appeal to our actual desires". Here, an attempt is made to justify hedonism in the following way: "The reason why it is always nonderivatively good for us to have pleasant experiences is that feeling pleasure is (always) intrinsically desired by us[7], and the reason why nothing besides pleasant experience is nonderivatively good for us is that we never desire anything else as an end, or for its own sake". (That is, an appeal is made to the object interpretation of the actual desire theory, regarded as a "method of justification").

Mill's argument

This is how Mill (1863) argued when he tried to justify evaluative hedonism in terms of psychological hedonism, i.e. the idea that pleasure alone is the object of our actual (intrinsic) desires. Mill's argument can be formulated in the following way:

(i) Psychological hedonism is true: nothing but pleasure (our own pleasure, that is) is desired as an end ("for its own sake", or "in and for itself").

(ii) The "justificatory" idea that "the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it" (ch. IV, p 32), or if we have value-for rather than value-period in mind: The idea that the sole evidence it is possible to produce that anything is nonderivatively good for people is that people actually desire it as an end.

(i) and (ii) suggest that hedonism is valid.

On Moore's (1903)[8] view, this is a bad argument, and the reason for this is that both (i) and (ii) are false. Is Moore right about this? Well, it seems quite clear that (i) is false: We often desire to feel pleasure, but it is (as Moore points out) certainly not the case that nothing but pleasure is (intrinsically) desired. (We could also add that people's actual intrinsic desires are different, and that a universal theory like hedonism can, for

---

[7]On the preference-hedonistic theory, it is (of course) analytically true that feeling pleasure is (always) intrinsically desired. However, the theory does not imply that it is good for us to feel pleasure *because* pleasant experience is intrinsically desired.
[8]Moore's criticism of Mill can be found in Moore (1903), pp 64-72.

this reason, not be justified in terms of what we actually desire).

If it is correctly understood, (ii) is not an implausible idea, however. It is of course true that the descriptive statement "we all desire to feel pleasure" does not *entail* the evaluative statement "it is good for all people to feel pleasure", but surely the truth of the former statement *suggests* (or indicates) that there is some truth in the latter statement. Or alternatively put, to believe that psychological hedonism is true while evaluative hedonism is false, this is a rather odd position.

In any case, even though Mill was mistaken in believing that we can appeal to our actual desires here, it seems that he understood something that Moore did not, namely that hedonism cannot be a justified theory of prudential value unless there is *something* in our "nature" that can "explain" why pleasant experience is the only thing that has nonderivative value for us. And if we cannot appeal to actual desire here, to what can we appeal instead? This is what we have to ask ourselves.

(2) The next question is whether it is possible to justify hedonism by appealing to our actual (intrinsic) evaluations. Now, it is not entirely clear what type of evaluations we are supposed to appeal to here, e.g., whether we should appeal to evaluations of the form "The only thing that is nonderivatively good for me is to feel pleasure" or to evaluations of the form "The only thing that it is nonderivatively good for all human beings is to feel pleasure". Suppose that we have the former type of "intrinsic" evaluation in mind. The idea that general value-for-statements like (H2) and (H3) can be justified in terms of people's evaluations can then be spelled out in the following way: "The reason why it is nonderivatively good for all people to have pleasant experiences is that every person believes that it is nonderivatively good for him or her to feel pleasure, and the reason why nothing besides pleasant experience is nonderivatively good for us is that every person believes that the only thing that is nonderivatively good for him or her is that he or she feels pleasure".

Is this a good subject-oriented argument for hedonism? I think not. In fact, this argument has the same weaknesses as "the appeal to our actual desires": it is simply not true that everyone believes that feeling pleasure is the only thing that is nonderivatively good for him or her, i.e. it is simply not the case that all people are hedonists with respect to what is nonderivatively good and bad for him or her. This means that

just as (H2) could not be justified by appealing to our actual desires, neither can it be justified by appealing to our actual evaluations[9]. Furthermore, we have different beliefs about what is good for us, and a universal theory like hedonism can only be justified by appealing to something which is common to everyone. What the hedonist has to show is that there is some *common* human feature that can explain why pleasant experience, and nothing but pleasant experience, has non-derivative value for us.

So, how can the hedonist meet this challenge? As far as I can see, there are only two routes open to him: He can either stop appealing to our actual desires or evaluations, and start appealing to our hypothetical desires or evaluations, or he can put certain restrictions on what an intrinsic desire or evaluation must be like in order to have "justificatory force". The latter route was taken by Wetterström (1986), when he tried to justify hedonism by appealing to our "H-evident subjective evaluations" (intuitions that are supposedly shared by all normal people). Wetterström's argument can be regarded as a "restricted version" of (2), a version that is based on the idea that our actual intrinsic evaluations do not have "justificatory force" unless they meet certain (rather strong) requirements.

Wetterström's argument

Let us first note that Wetterström is not really interested in finding out what types of facts that has nonderivative value for a person. Instead, he wants to find out what kinds of properties that "deserve to be appointed intrinsic v-characteristics", i.e. what properties that has intrinsic value-period. So, in order to make it clear how his ideas may be relevant to my purposes, I formulate his argument in my own terms.

Wetterström (1986) recognizes that an evaluative theory must be justified in a subject-oriented way: "How could any promising account of ethical justification fail to be based on certain assumptions about human nature", he writes (p 274). So, what are the "assumptions about human nature" that can, on his view, justify hedonism? His answer is (roughly) "the fact that all normal people share the same H-intuitions".

---

[9]And it is also worth pointing out that it is even harder to justify (H5) - the idea that the more pleasant a pleasant experience is, the better - in this way. We do not always believe that the more pleasant a pleasant experience is, the nonderivatively better it is for the subject to have it.

More specifically, he thinks that the reason why hedonism is valid is that "the H-intuitions of normal people are identical with hedonists' ultimate intuitions" (cf. ibid., p 267).

So, what does the term "H-intuition" refer to here? Well, let us first point out that all H-intuitions belong to the category of "subjectively intrinsic evaluations", i.e. that every H-intuition is (in Wetterström's terminology) of the form "X is good (or bad) as an end". In this context, however, it is more fruitful to regard H-intuitions and other intrinsic evaluations as being of the form "X is nonderivatively good (or bad) for P".

So, how do we determine whether a certain subjectively intrinsic evaluation is a H-intuition or not? This is where Wetterström introduces "the notion of an intuition being H-evident ['H' for 'Hume'] to a person": H-intuitions are "evaluations that are H-evident to actually existing people" (ibid., p 271), where the notion of H-evidence should be understood as follows: An evaluation *e* is H-evident to a person P (at a time t) if and only if as far as P is concerned, "you do something *absurd* or *impossible* if you question *e* - if you demand a reason for *e* or voice an objection to *e*" (ibid., p 271). He also adds that "[t]he labels of 'H-evident' and 'H-intuition' apply to concrete evaluations in the first place and to abstract evaluations in a derivative sense" (ibid., p 271). That is, if we have value-for-beliefs in mind, H-evidence is primarily attributed to evaluations of the form "a particular fact X is good (or bad) for a particular person P", where P is most probably *me*.

Now that we are familiar with Wetterström's terminology, his argument can be spelled out in the following way:

(i) An evaluation of the form "X is nonderivatively good (bad) for P" is valid if and only if it is H-evident to actually existing normal people[10]. That is, a person P's belief that a fact X is nonderivatively good for someone (e.g., for him) is valid if and only if P is normal and P would consider it absurd if someone were to question the belief. (It is important to note that this is a rather strong "principle of justification", e.g., it is much stronger than Mill's (ii)).

(ii) Most people have H-intuitions, and the H-intuitions of normal

[10]Or in Wetterström's own terminology: A property must be a v-characteristic in relation to a normal person's H-intuition in order to deserve to be appointed an intrinsic v-characteristic (cf. p 267).

people are identical with hedonists' ultimate intuitions (cf. ibid., p 267). The idea that the H-intuitions of normal people coincide with hedonists' ultimate intuitions can be divided into two parts: (a) Every normal person would (in fact) regard it as absurd if you were to question the idea that it is nonderivatively good for a person to feel pleasure (or if you would demand a reason for this idea); and (b) if someone believes that there are other things besides pleasant experiences that have final value for us (e.g., autonomous living), and if he thinks that it would be absurd or impossible to question this belief, then he is an abnormal person. The only "intrinsic evaluations" that it is absurd to question are the evaluations that constitute the hedonistic theory.

From (i) and (ii), we can conclude that hedonism is valid, or in Wetterström's terms, that "nothing but mental qualities deserve to be appointed intrinsic v-characteristics".

Let us now see whether this is a good argument or not, i.e. whether (i) and (ii) are plausible claims. Let us start with (ii), the factual premise:

On my view, (ii) is not an implausible idea, e.g., it is immensely more plausible than Mill's (i), i.e. psychological hedonism. Wetterström is (I think) "fundamentally right" in assuming that it is much harder for us to question the idea that it has final value for us to feel pleasure than it is to question the idea that it is nonderivatively good for us to have friends, or to engage in creative activity, or to live our lives in an autonomous way.

It seems that there are exceptions to both (ii:a) and (ii:b), however, particularly to the latter idea. This is rather obvious if we consider examples of a more holistic kind: Suppose that it is H-evident to me that it is bad to live in the experience machine, and that a life totally devoid of activity or intimacy is a bad life. Does this make me abnormal? I think not. So, it seems that it can not really be assumed that the H-intuitions of normal people are identical with hedonists' ultimate intuitions.

Let us now turn to (i), the "principle of justification". Is this a plausible principle? I think not (e.g., it is not as plausible as Mill's counterpart, i.e. (ii) on p 113). To say that an "intrinsic evaluation" should not be regarded as valid unless it is H-evident to normal people is a very strong requirement. But why should we accept such a strong requirement? In fact, it seems that the principle is designed with the purpose of justi-

fying hedonism. If it is harder for us to question (H3) than it is to question the idea that it is nonderivatively good for us to live autonomously, then we can save hedonism by stipulating that the "justificatory line" should be drawn somewhere in between the two.

In particular, it is not clear how Wetterström would handle the following objection: Suppose that I am "abnormal", and that my belief that it is nonderivatively good *for me* to have friends is H-evident to me. Why can't we conclude that it is (in this case) nonderivatively good for me to have friends? Do we really have to assume that the most plausible theory of prudential value is a universal theory, i.e. a theory which implies that if situations of a certain type are nonderivatively good for me, then they are also nonderivatively good for you, and vice versa? As far as I can see, Wetterström does not give us any reason why we should make this universalist assumption, so what is there to prevent us from regarding his justificatory principle as support for the following formal theory: "X is good for P if and only if (and because) it is H-evident to P that X is good for P"?[11] (A further problem with Wetterström's argument is that it is doubtful whether it can be used to justify the hedonist's comparative claims. For example, is it H-evident to all normal people that it is always better to have a more pleasant pleasure than to have a less pleasant pleasure?).

To conclude, Wetterström's argument is not a good-enough argument. This means that none of the attempts to justify hedonism by appealing to our actual desires or evaluations are successful. So what we have to ask ourselves now is this: Is it possible to justify hedonism by appealing to our hypothetical desires or evaluations, e.g., by appealing to what we would want if we were fully rational, fully informed, free of neuroses, assessing the matter "in a cool hour", or the like?

---

[11]Isn't this criticism a little "unfair"? After all, Wetterström had value-period in mind, and not value-for. Well, I think a similar kind of criticism can be formulated with regard to value-period, viz. the following one: Do we really have to assume that all intrinsic values are agent-neutral values? As far as I can see, Wetterström does not give us any reason why we should make this assumption, so what is there to prevent us from regarding his justificatory principle as support for the following theory: "P has a reason to promote X if and only if (and because) it is H-evident to P that X is intrinsically good-period"?

Appealing to hypothetical desires or evaluations

This is what such a justification would look like:

(3) In terms of desire: The reason why feeling pleasure is the only thing that is nonderivatively good for us is that if we were fully informed, free of neuroses, thinking clearly, and the like, this is the only thing we would desire as an end (That is, an appeal is made to the idealized desire theory, regarded as a "method of justification").

(4) Or in terms of evaluation: The reason why feeling pleasure is the only thing that is nonderivatively good for us is that this is the only type of fact to which a fully rational person P would attribute nonderivative goodness-for-P.

Are any of these two arguments a good subject-oriented argument for hedonism? I think not, for the following reasons:

First, they are both based on false factual assumptions: There is no reason to believe that if we were rational, we would all be hedonists. This claim is hard to prove, however.

Second, the principles of justification on which the arguments are based are not bad, but they both suffer from a certain kind of incompleteness. Or alternatively put, as they stand, these principles are not really subject-oriented at all. It can hardly be doubted that these principles have a certain "justificatory power", e.g., if I honestly came to believe that I would be a hedonist if I were rational, this would at least suggest that hedonism is a plausible theory. Now, this is the reason why the principles are incomplete: Suppose that if we were rational, we would all be hedonists. In a sense, this would be a fact about us, but it is not a fact that we would rest content with in this justificatory context; we would still want to know *why* we would be hedonists if we were rational. On the assumption that pleasant experience is the only thing a rational person would desire as an end, we would need to know what it is about us human beings that explains this alleged fact. If we were rational, we would have a clearer and more unclouded view, but what is it (e.g., in us, or about us) that we would see if we were rational that we do not see now? For example, is pleasant experience the only thing we *really* desire, deep down? Is some version of psychological hedonism true after all? (Are all our "non-hedonistic" intrinsic desires and evaluations somehow "mistaken"?) How would one justify such a claim? Do these deep desires exist at all?

In short, the second (more general) reason why (3) and (4) will not

do is that a subject-oriented justification of hedonism cannot be complete unless it refers to what the subject is *actually* like, unless it "takes hold" of something that is actually there, "in" the subject. But if (1) and (2) do not work either, i.e. if the hedonist cannot appeal to our actual desires or evaluations, then what human feature (if any) is left for him to "take hold of"? It seems that there is nothing there. In short, it seems that all attempts to justify hedonism in a subject-oriented way are failures.

This ends the list of pro-arguments. Let us now look at how the hedonistic theory can be criticized.

3.2. Arguments against hedonism

Before we take a closer look at some of the counter-arguments that can be directed against hedonism, let us first make a few general remarks about what these arguments are (in fact) like, and what they have to be like in order to be successful. What would it take to refute the hedonistic theory?

Arguing against (criticizing, refuting) hedonism: some general remarks

Let us first point out that it is not just a complete justification of hedonism that must be subject-oriented; a complete *refutation* of the theory must also be based on some kind of conception of human nature: the anti-hedonist must (ideally) give us an account of human nature that "explains" why hedonism is a bad theory of prudential value.

This is most probably *one* reason why some anti-hedonists (e.g., Nozick) tend to point out that *we* (most of us) intrinsically want other things besides pleasant experiences, that we are not really hedonists (that we tend to reject (H2)), and so on[12]. Now, it is important that these references to our desires and evaluations are recognized for what they are, viz. as a kind of subject-oriented criticism of the theory. (If this is not understood, these references will simply be dismissed as irrelevant, which is not very fair). Furthermore, it is also important that

[12]The other reason is rhetorical: To appeal explicitly to *our* intuitions, to what *we* want and believe (e.g., as in "we all know that..."), can be an effective way to convince an audience.

these subject-oriented arguments against hedonism are properly understood, viz. as a kind of challenge to the hedonist: "The thing in us that makes hedonism plausible is not our actual desires or evaluations, so you'd better tell us what it is!". If this is so, all the references to the fact that our desires and/or evaluations are often "anti-hedonistic" are best regarded as counter-pro-arguments, i.e. as arguments against possible arguments for hedonism (e.g., like Mill's argument). That is, what these arguments do is that they make it hard (or even impossible) for the hedonist to come up with an acceptable subject-oriented justification of his theory.

Most arguments against hedonism are not subject-oriented, however, but (rather) *attempts to show that hedonism has unacceptable implications*, i.e. that the theory has implications which are inconsistent with certain valid (or widely accepted) judgements. Or alternatively put, the idea is to demonstrate that if the hedonistic theory is tested against other (more particular or more specific) value judgements, it will not survive the test.

The value judgements against which the theory is tested are almost always more specific than hedonism itself, but they rarely include references to particular persons. This gives rise to the following problem: I have claimed (e.g., in section 1.3) that we should (in contexts of justification) give much more weight to our judgements about particular cases than to our more general judgements, e.g., if any of the particular statements that can be deduced from a general theory are inconsistent with our judgements about particular cases, we should tend to give up the theory[13]. But what if the value judgement with which the theory is inconsistent is on the same level of generality, as in the following case: "Hedonism implies that "what you don't know can't hurt you", e.g., that it can never be nonderivatively bad for a person to be deceived without knowing it. But it *is* sometimes bad for a person to be deceived without knowing it, and this means that hedonism is wrong". Here, the

[13]The reason why we should not *always* choose to give up the theory is this: We think that our judgements about particular cases should be granted a privileged position (that we should try to "save" our intuitions about particular cases), but we also want them to be coherent (consistent) with each other. And it seems that the only way in which such coherence can be assured is *via* some general theory (cf. section 1.3 above); this is why it is important that we have some sort of theoretical ambition in this area. And if we have this ambition, it seems that we can't *always* give more weight to our particular judgements.

following question arises: Is it really plausible to give more weight to the idea that it is sometimes bad for a person to be deceived without knowing it than to hedonism itself? Is there any reason why we should grant such general judgements a more "privileged position" than our general theories? Well, I think there is, but only if the judgement in question is (so to speak) more specific than the theory.

This problem is closely connected to another set of problems, viz.: How should arguments of this kind (arguments that purport to show that hedonism has unacceptable implications) be understood? What do they purport to show, what do they actually show, and how should they be assessed?

To the extent that it is possible, we will regard all arguments of this kind as "objective reasons" against hedonism, i.e. as attempts to refute the theory. Now, this suggests that we should evaluate these arguments as follows: First, we should ask ourselves whether the evaluation to which the argument appeals (e.g., the idea that it is sometimes bad for a person to be deceived without knowing it, or that a life in the experience machine would be a bad life) is valid or justified, and second, we should ask ourselves whether the evaluation really constitutes a threat against hedonism (whether it is relevant), i.e. we should ask "*if* the evaluation is valid, does this mean that hedonism has to be rejected?".

Now, it seems that the hedonist and the anti-hedonist rarely disagree on issues of relevance, i.e. they tend to agree on what the implications of the theory are, and what would constitute a threat against the theory. There are often disagreements on whether the evaluation to which a certain argument appeals is justified, however, and this gives rise to the following central question: Is it at all possible to determine whether such an evaluation is justified or not, and if it is, how should it be determined?

Roughly speaking, a certain crucial anti-hedonistic evaluation is justified if and only if it is consistent with other evaluations *and* if it has "inductive support" from those of these evaluations which are more specific (or more particular) (cf. section 1.3 above). That is, I strongly reject the generalist-deductivist idea that a more specific evaluation (intuition) is justified only if it can be deduced from a valid (or justified) general theory. This suggests that we should not think of a more specific anti-hedonistic evaluation as *presupposing* that some alternative

general theory is valid or justified. It is probably true that if a certain evaluation can be used against hedonism, then it *can* also be deduced from some alternative general theory, but this does not imply that the anti-hedonistic evaluation cannot be valid or justified *unless* the general theory from which it can be deduced is valid. For example, the idea that a life in the experience machine would be a bad life does not presuppose that some alternative general theory is true, i.e. *if* it happens to be valid, the reason for this is *not* (necessarily) that it can be deduced from a valid general theory. The validity of a more specific anti-hedonistic evaluation is (to a considerable extent) independent of the validity of all alternative general theories, and this is why it might give "inductive support" to some non-hedonistic general theory, like the desire theory. This is how we should view the fact that the crucial anti-hedonistic intuitions can often (or always) be deduced from some alternative theory, viz. as a fact that explains why most arguments against hedonism are also arguments (inductive reasons) for some other theory.

Now, we all know that there is a lot of intractable disagreement in this area, and this might give rise to the suspicion that it is often impossible to determine whether arguments of this kind (arguments that purport to show that hedonism has unacceptable implications) are good or not. So maybe we should regard them "subjectively" rather than "objectively" (cf. pp 106-107), i.e. as more or less successful attempts to convince people that hedonism is a bad theory. The problem with this approach is that it makes it impossible to evaluate the arguments in a general way. For example, we can not say that an argument has to convince (in a rational way) everyone (or every normal person) in order to count as a good subjective reason. The reason for this is that arguments of this type will only convince those people who share the relevant intuitions, like anti-hedonists or certain tender-minded or unreflected "hedonists", i.e. they will not convince a reflected, tough-minded hedonist about anything (if someone accepts the implications, there is little or nothing that can be done to convince him in a rational manner). In short, if an argument is regarded subjectively, it can only be evaluated in relation to some person (audience) or other, a person that is already equipped with a number of beliefs, evaluations, desires, and so on, and this makes the subjective approach rather unattractive. So, it seems that the only reasonable thing to do at this point is to present the arguments and let the reader decide for himself whether

they are good or not.

The arguments against hedonism: A brief overview

Before we look at the counter-arguments themselves, it is important to notice that these arguments have somewhat different "targets", and it might be helpful to classify them on basis of what version or versions of the hedonistic theory that they purport to refute (or succeeds in refuting), or alternatively, on basis of what hedonistic claim or claims they purport to disprove. If the arguments are classified in this way, they will fall into three groups:

First, there are the arguments which are directed at all kinds of hedonism, i.e. at (H1) (the Experience Requirement) and/or (H2)[14]. This is where we find the arguments which purport to show that a person's well-being can be directly affected by things he doesn't know anything about.

In the second group, we find a number of arguments that are primarily directed at (H6): arguments that purport to show that experience is not the only thing that matters, that the value that a life has for the person who is living is does not just depend on how it feels from the inside. Now, the idea that there are other things besides experience that matters is not very precise, so we must ask ourselves: How (in what way) do these other things matter? Or alternatively put, what other hedonistic claims besides (H6) do these arguments (if they are valid) hit? Well, they hit (H4)-(H5), and maybe also (H3), but not necessarily (H1) and (H2). That is, they hit pure hedonism, but they need not hit modified hedonism. This should not surprise us; one of these arguments (viz. Mill's "pig-argument") is constructed by a modified hedonist, and there is really no essential difference between this argument and arguments like Nozick's experience machine[15].

Arguments of the third kind all purport to criticize pleasures of cer-

---

[14]Here, it is important to notice that only some of the arguments are of this type. That is, most of the arguments against hedonism are not really directed against hedonism "as such".

[15]However, if we take a closer look at the arguments that have been put in this group, we see that there is (in some cases) more to these arguments than the attack on (H6), e.g., several of these arguments contain a subject-oriented element, and some of the things that Griffin says seems to be directed against (H1).

tain kinds. These arguments are directed against pure hedonism, and they are delivered from the standpoint of some kind of modified hedonism. That is, (H1) and (H2) are never put in question by these arguments, and it is therefore possible (but not necessary) to regard the arguments as parts of a "discussion" that is (so to speak) internal to hedonism. There are two kinds of arguments in this group, viz. (i) arguments against (H3), the thesis of Unrestrictedness, and for (RH3), some restriction claim, and (ii) arguments against (H4)-(H5) (and indirectly against (H6)) and for (MH4), some claim concerning relative weights.

These arguments may well be "extensionally equivalent" with the arguments in the second group (they may well hit the same hedonistic theses), and the objections in the third group are (moreover) often "derived" from the objections in the second group. However, there are still good reasons for keeping the two groups separate. First, the arguments in the second group are often aimed at hedonism as such, but this does not hold for the arguments in the third group. Second, the arguments in the second group are first and foremost directed against (H6), and only indirectly against (H4)-(H5), while the arguments in the third group are of a more "atomistic" (or "local") kind; they are primarily directed against (H3) and/or (H4)-(H5), and only indirectly against (H6).

Let us now look at the counter-arguments themselves. We will start with the arguments that belong to the first group.

Arguments directed against (H1) and/or (H2)

Arguments of this kind purport to show that there are other things besides pleasure and suffering that are nonderivatively good and bad for a person, or more specifically, that a situation need not have any "experiential content" at all in order to have nonderivative value for a person. There are at least two kinds of arguments for the idea that non-experiential situations can have nonderivative value for us, viz. (i) arguments which appeal to the idea that a person's well-being can be directly affected by things he does not know anything about, and (ii) arguments that purport to show that some non-experiential version of the desire theory (e.g., the Success Theory) is superior to hedonism. It is worth noting that arguments of these kinds are often, but not

always, subject-oriented arguments that appeal to our actual desires or evaluations.

Arguments which purport to show that a person's welfare can be directly affected by things he doesn't know anything about

The hedonist claims (roughly) that a situation cannot benefit a person unless it gives him pleasure, and it cannot harm him unless it makes him suffer. Now, this view does not really imply that "what you don't know can't hurt you" (a person may, after all, be exposed to nuclear radiation without knowing it), but it does imply that a person cannot be benefited or harmed "directly" by what he doesn't know: If a person does not know that a certain situation X holds, and if the occurrence of X does not affect P's experiential life negatively in other ways, then P can not be harmed by X[16]. This means that if there are cases where a person can be harmed directly without knowing it, the hedonistic theory must be false. So, are there any cases of this kind, and if there are, what would they be like? The following examples are offered by Nagel (1970):

> [On the hedonistic view] /.../ what you don't know can't hurt you. It means that even if a man is betrayed by his friends, ridiculed behind his back, and despised by people who treat him politely to his face, none of it can be counted as a misfortune for him as long as he does not suffer as a result (p 4).
>
> Loss, betrayal, deception, and ridicule are on this [hedonistic] view bad because people suffer when they learn of them. But it should be asked how our ideas of human value would have to be constituted to accommodate these cases directly instead. One advantage of such an account might be that it would enable us to explain *why* the discovery of these misfortunes causes suffering - in a way that makes it reasonable. For the natural view is that the discovery of betrayal makes us unhappy because it is bad to be betrayed - not that betrayal is bad because its discovery makes us unhappy (ibid., p 5).

Now, it is important to see that this argument does not presuppose any particular view on *why* is it nonderivatively bad for a person to be

---

[16]True, it is possible that hedonism can allow for subnoticeable changes in well-being, but I regard this idea as somewhat peripheral in this context.

betrayed, deceived, or ridiculed behind his back, e.g., whether it is bad because the person happens to have an intrinsic aversion against it (as the desire theorist who accepts the object interpretation would say), or because it is simply bad for him (as an objective list theorist might say). That is, as the argument stands, we only know that it presents a serious threat to hedonism; we don't know what other theory that is supported by it.

A special case: Post-mortem events

A rather special case is this: Suppose that a person has intrinsic desires about what will happen after his death, e.g., that his money will be spent to relieve famine, and that no organs will be removed from his body. He then dies, but his wishes are not respected. His relatives use the money to buy drugs, and medical students dissect his body. Is the person being harmed here? Is it plausible to claim that a person's well-being can be affected in a negative way, after his death? The hedonist does not think so, e.g., on his view,

> a man is not injured if his wishes are ignored by the executor of his will, or if, after his death, the belief becomes current that all the literary works on which his fame rests were really written by his brother, who died in Mexico at the age of 28 (Nagel (1970), p 4).

If we don't share this intuition, we have to reject hedonism. An example of a philosopher who disagrees with the hedonist on this point is Aristotle. In *NE*, book I, chapter 10, he "writes" that

> /.../ both evil and good are thought to exist for a dead man, as much as for one who is alive but not aware of them; e.g. honours and dishonours and the good or bad fortunes of children, and in general of descendants (p 19).

As far as I can see, there are two possible reasons for sharing this Aristotelian view: (i) One may appeal to the preferences that the person had when he was still alive (this move is not open to an objective list theorist like Aristotle, however). (ii) One may also appeal to the idea that since we are creatures who view ourselves both "from within" and "from without", an assessment of a person's well-being should not be made from a first-person-perspective only: it is also necessary to adopt a third-person-perspective (This is probably the kind of reasoning on

which philosophers like Aristotle and Nagel base their views).

Is this a good argument? I think not[17].

Arguments that purport to show that the desire theory is superior to hedonism

As I see it, there are at least two arguments of this type. The first one is rather general and theoretical, while the other argument purports to show that the desire theorist can explain something that the hedonist can not, viz. that death is an evil for the person who dies.

A general argument against the experience requirement

This argument purports to show that some non-experiential version of the desire-fulfilment theory (e.g., the Success Theory) is superior to hedonism. The argument has two steps:

(a) The first step consists in trying to establish that Preference-Hedonism and/or the experience-oriented Success Theory is more plausible than Quality Hedonism. This step was taken on pp 100-104 above, i.e. it seems that we can assume that if E1 and E2 are two experiences, and if the experiencing subject has an intrinsic now-for-now preference for E2 over E1, then it is nonderivatively better for the subject to have E2 than to have E1.

(b) We have now entered the framework of the desire theory. That is, we have (so to speak) left hedonism behind, but we have not yet abandoned the experience requirement. The second step of the argument consists in showing that the ordinary Success Theory is more plausible than the experience-oriented version of the theory, i.e. that the experience requirement (i.e. (H1)) is implausible *within the framework of the desire theory*[18]. This is an example of how this could be established: Suppose that P intrinsically wants to have self-knowledge (to have true beliefs about himself), and that this desire is stronger than his (intrinsic) desire to avoid certain kinds of suffering. Now, suppose that there is a

---

[17]It is worth noting that this is (on the assumption that death is annihilation) naturally explained by an idea put forward in appendix C, viz. that all value-for-P is value-for-P-at-*t*. It may even provide some support for the idea.

[18]That is, the second step of the argument is based on the assumption that (H2) is already refuted, and it is (therefore) not really necessary to include it in this context. The reason why I include it anyway is that it shows that *if* (H2) is implausible for a certain reason (i.e. because the experience-oriented Success Theory is superior to hedonism), *then* we must also regard (H1) as implausible.

certain fact X about P (or his history) which would, if P found out about it, make him suffer (if P would believe that X, this would produce an experience in him that he would desire not to have). But all the same, P happens to prefer painful knowing to pleasant not-knowing (or "blissful ignorance"). In a case like this, it would be better for P to find out about X than to remain in an ignorant state, even if he, "experientially speaking", would prefer the latter state.

This point can also be formulated in a more general way: If we can appeal to preferences within the domain of experience, why can't we appeal to preferences "all the way"? If our preferences do (in fact) matter in this context, what reason is there for disregarding the fact that "we do seem to desire things other than states of mind, even independently of the states of mind that they produce" (cf. Griffin (1986), p 9), and what reason is there for accepting the idea that "when, with eyes open, I prefer something not a mental state to a mental state and so seem to value the former more than the latter", it is in fact better for me to have what I value less (cf. ibid., p 10)? In short, it seems that the experience requirement is quite arbitrary within the framework of the desire-fulfilment theory.

Is death an evil for the person who dies?

The next argument against the hedonistic theory is based on two assumptions, viz. (i) the idea that death can be an evil for the person who dies, and (ii) the idea that a theory of prudential value cannot be plausible unless it is compatible with (i), and unless it can explain (i). The main thing that the argument purports to show is that hedonism is incompatible with the idea that death can be an evil for the person who dies, i.e. *if* (i) and (ii) are plausible assumptions, then hedonism must be regarded as an implausible theory[19]. However, it also purports to show that there is at least one theory of prudential value that is compatible with (i), and that can explain (i), viz. the object interpretation of the desire theory. (The reason why this is necessary is this: If no theory of prudential value were compatible with (i), (ii) would be a strange condition).

Let us first try to show that if hedonism is correct, then a person's

[19]Hedonism *qua* conception of nonderivative value-for, that is. Hedonism *qua* theory of final value-period is certainly compatible with the idea that it is bad-period that a person dies!

death can not really be regarded as bad for this person.

On the hedonistic theory, the only facts that can have nonderivative value for a person P are facts of the type "P has an experience E". If we combine this idea with the idea that the subject for which a particular situation is good or bad is always a-person-at-a-time, and that we should always think of value-for-P as value-for-P-at-*t* or value-for-P-at-any-*t* (but never as value-for-P-over-time), we get the following idea: The only facts that can have nonderivative value for P-at-*t* are facts of the type "P has an experience E at *t*". This idea is an instance of a more general idea, viz. the idea that a situation can not have nonderivative value for P-at-*t* unless it occurs at *t* (cf. appendix C).

The hedonistic conception of *derivative* value-for can be characterized as follows: A situation is derivatively good for a person if it gives a causal contribution to him having certain pleasant experiences in the future, or if it makes certain future pleasures possible, or if it makes certain future sufferings impossible or improbable, or the like. A situation is derivatively bad for a person if it gives a causal contribution to future sufferings, or if it makes certain future sufferings possible, or if it makes certain future pleasures impossible or improbable, or the like. This means that if a certain situation has derivative value vis-a-vis a certain nonderivatively valuable situation, then the derivatively valuable situation can not occur after the nonderivatively valuable situation.

Let us now assume (as Epicurus and Lucretius did) that death is real, i.e. that death is annihilation. This means that a person's death is not an experience, and it can therefore (on the hedonistic theory) not be nonderivatively bad for this person. So, can death be derivatively bad for the person who dies? Well, a person's death can not be experienced by himself, and it can not cause any experiences in him, and it can (therefore) not be derivatively bad for him in any of these two ways. Does this show that a hedonist must accept the Epicurean idea that a person's death is not an evil for this person[20]? Maybe not. Maybe the idea that death is an evil for the person who dies is compatible with hedonism after all, viz. for the following reason: "It is bad for most of us that we are going to die because our deaths will deprive us of all the future pleasures we would have if we would not die. And for most

[20]It is (of course) assumed here that if X is bad for P, then it is either nonderivatively bad for P or derivatively bad for P. Can this assumption be questioned?

dead people, it was bad for them that they died, because if they had not died, they would have continued to have pleasant experiences".

This idea is based on a total misunderstanding of what annihilation means, however. A person's life is not a piece of property which belongs to him (it is not something that he can lose in any literal sense). Neither is it a state that he can either be in or not be in. To lose one's life is to stop existing, and it is therefore inaccurate to conceive of a person's death as a loss *for him*; to "lose one's life" is to disappear, and when a person dies, there is no longer anyone there who can meaningfully be said to have lost a life. The same kind of reasoning can be applied to the issue of deprivation of future experiences: To die is not to be deprived of future experiences, because there is no one there anymore who can be regarded as deprived. To die is not like losing one's fortune or one's position; in the latter cases the person continues to exist, i.e. there is someone around which can be identified as the deprived one.

Or alternatively put, there is no time *t* such that P's death is bad for P-at-*t*. A future event can not be bad (nonderivatively or derivatively) for P-now; P's death cannot be bad for P-when-it-occurs (since it is neither an experience nor an experienced event); and P's death can not be bad for P-after-death, because if death is annihilation, there is no such thing as P-after-death. In short, Epicurus was right; if one assumes that hedonism is correct, and that death is annihilation, then it follows that death can not be an evil for the person who dies. So if you think that your own death is (or will be) an evil (for you), you cannot be a Hedonist.

However, the desire theory is compatible with the idea that a person's death can be an evil for this person, and it can also explain this idea. The reason why it is bad for most of us have a very strong intrinsic aversion against dying. We do not want to die, period.

Is this a good argument? I think not, and the main reason for this is that (ii) is a faulty premise. On my view, there is no plausible theory of value-for that can *explain* the alleged fact that death can be an evil for the person who dies (which of course does not prevent an objective list theorist from simply *claiming* that it is nonderivatively bad for us to die[21]). For example, a desire theory can only be consistent with the idea

[21]There is of course also the possibility that P's death is not really an evil for P

that it is bad for a person to die if it assumes that it can be good for a person to have his prospective desires fulfilled, but this is not a plausible assumption (I will try to show this in section 4.3). A "synchronistic" desire theory can explain why it is good for a person to continue living[22], however, but so can the hedonist, i.e. there is really no difference between the theories in this particular respect.

Arguments that purport to show that pleasure is not all that matters

Can pleasant sensation ever be sufficient to make a life good? Nagel's contented infant

The first argument is not an attempt to show directly that pleasure is not he only thing that matters; what it purports to show is rather that pleasant sensation is not sufficient to make a human life good.

The pure hedonist claims that the value that a certain life has for the person who is living it is a function of how much pleasure and suffering it contains. The more pleasure and the less suffering a life contains, the better this life is for the person who is living it, period. Here, it does not matter at all to what category a person's pleasures belong, e.g., whether they are sensations or emotions; all that matters is how pleasant these pleasures are. This means that if a person has a lot of pleasant sensations (and if he does not suffer too much), this is (on pure hedonism) sufficient to make his life good, no matter what other features this life has. On this view, a life that is totally devoid of emotional content might well be good for the person who lives it, and so might an extremely fragmentary life (a life that consists in a sequence of disconnected pleasant sensations). The theory also implies that "the life of a pig" might well be better for a person than "the life of a human being", i.e. that it does not really matter (from an evaluative point of view) whether a life is "recognizably human" or not. These implications are (according to the argument) not acceptable, however, and pure hedonism therefore has to be rejected.

---

after all, i.e. (i) may well be a faulty premise too! Cf. appendix C.

[22]Notice that the idea that it is good for a person to continue living (i.e. not to die) does *not* imply that it is bad for a person to die, or that it is better for a person to continue living than to die. However, it strongly suggests that it is bad-period that someone dies, and that it is better-period that a person continues living than that he dies.

Let us now look at a "more detailed" version of this argument. The argument is based on an example given by Nagel (1970):

> Suppose an intelligent person receives a brain injury that reduces him to the mental condition of a contented infant, and that such desires as remain to him can be satisfied by a custodian, so that he is free from care. Such a development would be widely regarded as a severe misfortune, not only for his friends and relations, or for society, but also, and primarily, for the person himself (p 5)[23].

The pure hedonist would not necessarily regard the injury as a "misfortune for the person himself", however. That is, on his view, it is not necessarily bad for a person to be "reduced to the mental condition of a contented infant"; it all depends on whether there happens to be more pleasure in his life after the injury or not. This is what a pure hedonist might say:

> If we did not pity him then [when he was three months old], why pity him now; in any case, who is there to pity? The intelligent adult has disappeared, and for a creature like the one before us, happiness consists in a full stomach and a dry diaper (ibid., p 6).

This is not an acceptable view, however, and pure hedonism should therefore be rejected. But again, if there is a tough-minded hedonist who accepts this implication, there is really nothing we can do to convince him[24].

---

[23]He then adds that it is the intelligent adult who has been *reduced* to the condition of a contented infant, and not the "contented infant", who should be regarded as the subject of the misfortune (ibid., pp 5-6). Now, if we share this intuition, the argument really belongs in the same category as the "death-is-an-evil-for-the-person-who-dies"-argument above. So, let us therefore (for the sake of argument) assume that the "contented infant" and the intelligent adult are the same person, and that it is this person's well-being who is affected (for better or worse) as a result of the injury.

[24]Before we move on to the next argument, it should be noticed that Mill's (1863) idea that "[i]t is better [for a human being, that is] to be a human being dissatisfied than a pig satisfied; better to be a Socrates dissatisfied than a fool satisfied" (ch. II, p 9) can also be regarded as a version of this argument, but only if it can be assumed that a pig's (etc.) level of satisfaction is dependent on one thing only, viz. on what how much pleasant *sensation* there is in his life. We will return to Mill's argument below.

Let us now look at a number of arguments which purport to show that experience is not all that matters, or more specifically, that the value of a life (for the person who is living it) does not just depend on what its experiential content is like, i.e. on how much pleasure and suffering this life contains. This is not all that these arguments purport to show, however; some of them can also be viewed as arguments for the stronger thesis that a person's well-being can be affected by things he does not know anything about (cf. pp 126-128 above), or for the idea that it is better for a person to have pleasures that are based on true beliefs than to have pleasures that are based on false (or illusory) beliefs (cf. (RW2) on p 147 below).

It is worth noting that all these arguments appeal to the fact that hedonism is incompatible with our "ordinary notion of well-being", that we (most of us) regard other things besides pleasure as nonderivatively valuable, that we often take pleasure in something because we regard it as nonderivatively valuable (rather than the other way around), and so on. Or alternatively put, the arguments are (so to speak) based on the idea that it is proper to adopt an immanent perspective in this context, and the hedonist is accused of disregarding our own perspectives[25]. This does not mean that the arguments are based on the assumption that the desire theory is correct, however. An appeal is made to the fact that most of us are (in fact) *not* hedonists, and the reason for this is that it is assumed that there is an intimate connection between what we regard as good and bad for us and what is (in fact) good and bad for us. Now, it is true that this assumption is compatible with the desire theory, but it is also compatible with the objective list theory. What these two theories disagree upon is not whether there is an intimate connection or not between our evaluations and what is (in fact) valuable, but on the nature of this connection. On the object interpretation of the desire theory, the connection is conceptual, and the hedonist is wrong because he disregards what we want and do not want, period. The non-internalist pluralist, on the other hand, has a different idea about *why* our actual evaluations should not be disregarded. On his view, there are things which are good and bad for us regardless of what our desires and aversions are, and he therefore conceives of the

[25]Besides hedonism, there is another theory that also tends to adopt a "transcendent perspective", viz. the satisfaction interpretation of the desire theory. Cf. Rabinowicz and Österberg (1996), pp 12-13. Cf. also section 5.3 below.

connection between evaluations and facts of value as contingent; the reason why our evaluations are often correct is that we (most of us, or some of us) have an ability to *recognize* (or *discover*) what is, in fact, nonderivatively good and bad for us. On this view, the hedonist's mistake consists in not recognizing that "we" are often right about what is good and bad for us. However, it should be noted that the arguments that are to follow are not just compatible with the objective list theory; some of them (e.g., the first three) also appeal directly to what they take to be "objective prudential values". Let us now look at the arguments themselves:

Griffin's argument

This is how Griffin (1986) criticizes the hedonistic theory. It is mainly directed against the pure version, but if it is successful, it also hits all modified versions.

> It [the Experience Requirement] seems in the end simply too drastic. It bans things that our ordinary notion of well-being cannot, without damage, do without. It is common that, as many persons' values mature, such things as accomplishment and close authentic personal relationships come more and more to fill the centre of their lives. If the Experience Requirement excludes these values from 'utility' [i.e. if it tells us that these things are not nonderivatively good for us], then 'utility' will have less and less to do with what these persons see as making their own lives good. And those values do seem excluded. Suppose that someone is duped into thinking that those close to him are behaving authentically. What enters experience is the same whether he has the real thing or a successful deceit. But it is only the real thing, he thinks, that makes his life better. According to the enjoyment account [hedonism], what affects well-being can only be what enters experience, and the trouble is that some of the things that persons value greatly do not. My truly having close and authentic personal relations is not the kind of thing that can enter my experience; all that can enter is what is common to both my truly having such relations and my merely believing that I do. And this seems to distort the nature of these values. If I want to accomplish something with my life, it is not that I want to have a *sense* of accomplishment. This is also desirable, but it is different from, and less important than,

the first desire. /.../ If either I could accomplish something with my life but not know it, or believe that I had but not really have, I should prefer the first. That would be, for me, the more valuable life (p 19)[26].

Is this a good argument? Yes, it is.

Rachels' argument

This is Rachels' (1986) story about Wonmug, followed by his own comments:

> Wonmug was a somewhat stupid but very vain college student interested in physics. The other students would amuse themselves by making fun of him; but to his face they pretended to have great respect for his intellect. As a result, Wonmug came to believe himself to be their intellectual leader. The others thought this very funny.
>
> /.../ Gradually, the entire scientific establishment came to participate in the charade. /.../ Finally, he was awarded the Nobel Prize. As Wonmug delivered his pompous acceptance speech, the woman he loved, but who ridiculed him behind his back, sat beside him beaming with false pride, and the members of the Swedish Academy could barely keep from laughing. When Wonmug died he sincerely believed himself to have been the greatest and most beloved figure in science since Einstein.
>
> /.../ Was Wonmug a fortunate individual, or an unfortunate one? Did he have a good life, or not? In an obvious way he was very fortunate. He received honours that most of us can only dream about. But of course there was something radically wrong. Although he *thought* he achieved great things in science, he really achieved nothing. Although he *thought* he had many friends, he really had no friends. But what is wrong with that? Hedonism says that all the things we value - knowledge, achievement, the love and respect of other people - are good *only in that they cause pleasant states of consciousness*. It is only the states of consciousness that are good 'in themselves'. Wonmug, in fact, had all the states of consciousness

[26]To see why there is (most probably) an appeal to objective values here, imagine a person who evaluates (in a positive way) degradation and irrational behaviour rather than achievement and friendship. Would Griffin consider it to be better for this person to be "authentically degraded" than to just have a "sense of degradation"? I think not.

associated with achievement, respect, and the rest. So according to hedonism he had everything good associated with these things. His life was just as good as Einstein's, and maybe even better.

/.../ [Wonmug] is happy only because he is ignorant of what is going on. And if he were to discover what is really happening, he would /.../ come to see himself /.../ [as a pathetic figure], and his happiness would be shattered. But why would his happiness be diminished? The answer is that what *he* values (and what we value too) is such things as achievement and friendship. It is because he values these things, and because he thinks he has them, that he is happy. /.../

This is /.../ the 'logical mistake'. In saying that achievement and friendship are good because they make us happy, hedonism gets things the wrong way round. They are not good because they make us happy. Rather, having them (and other things like them) makes us happy because we recognize them *as goods*[27]. To explain their value, then, we have to look elsewhere than to the conscious states that accompany them (pp 46-48).

This is also a good argument.

Nozick's "Experience Machine"

Nozick (1974) and (1989) asks us to imagine an experience machine that could give us any experience that we desired. Being plugged into the machine, one would all the time be floating in a tank, with electrodes attached to one's brain, and the machine would stimulate the brain in any way pre-programmed. While in the tank, one would not know that one is there; one would think that it is all actually happening. Nozick now asks us (rhetorically) whether we would plug into this machine for life. Well, for most of us, the answer is "no", and the reason for this is (of course) that "something matters to us in addition to experience" (Nozick (1974), p 44), or alternatively, that we "care about things in addition to how our lives feel to us from the inside" (Nozick (1989), p 104).

---

[27]That is, the reason why friendship makes us happy is (among other things) that we value it, but the reason why it is good for us is (on Rachels' view) not that we value it. On the contrary, the fact that it is objectively good for us to have friends might explain why we value it. That is, Rachels' argument does not appeal to our preferences, but to objective prudential values.

So, what is it that matters to us in addition to experience? What is it that we want (value, care about, etc.) that the experience machine can not give us? This is how Nozick himself answers this question:

> First, we want to *do* certain things, and not just have the experience of doing them. /.../ A second reason for not plugging in is that we want to *be* a certain way, to be a certain sort of person. Someone floating in a tank is an indeterminate blob. /.../ Thirdly, plugging into an experience machine limits us to a man-made reality, to a world no deeper or more important than that which people can construct. There is no *actual* contact with any deeper reality, though the experience of it can be simulated. Many people desire to leave themselves open to such contact and to a plumbing of deeper significance (Nozick (1974), p 43).

> What is most disturbing about them [e.g., experience machines] is their living of our lives for us. /.../ Perhaps what we desire is to live (an active verb) ourselves, in contact with reality. (And this, machines cannot do *for* us) (ibid., pp 44-45).

> We care about more than just how things feel to us from the inside; there is more to life than feeling happy. We care about what is actually the case. We want certain situations we value, prize, and think important to actually hold and be so. /.../ We want to be importantly connected to reality, not to live in a delusion. /.../ What we want and value is an actual connection with reality. /.../ To focus on external reality, with your beliefs, evaluations, and emotions, is valuable *in itself*, not just as a means to more pleasure or happiness. /.../ We do not, of course, simply want contact with reality; we want contact of certain kinds: exploring reality and responding, altering it and creating new actuality ourselves (Nozick (1989), p 106).

> No doubt, too, we want a connection to actuality that we also share with other people. One of the distressing things about the experience machine, as described, is that you are alone in your particular illusion (ibid., p 107).

> It seems too that once on the machine a person would not make any choices, and certainly would not choose anything *freely*. One portion of what we want to be actual is our actually (and freely) choosing, not merely the appearance of that (ibid., p 108).

Now, if these claims are understood as empirical claims about how people actually are, this is not a good argument against hedonism: From the empirical fact that we want other things besides pleasant experiences, we can not conclude that it would not be good for us to spend our lives in the experience machine. However, there is no reason to believe that Nozick's claims should be understood in this way (only). They should, rather, be understood as substantive evaluative claims about what is good and bad for people. Nozick himself is quite clear about this:

> Notice that I am not saying simply that since we desire connection to actuality the experience machine is defective because it does not give us whatever we desire /.../. Rather, I am saying that the connection to actuality is important whether or not we desire it - that is *why* we desire it - and the experience machine is inadequate because it doesn't give us *that* (ibid., pp 106-107).

That is, the fact that he makes almost all his points in terms of what matters to us, what we want, what we care about, what we value, and what is important to us, is best regarded as a kind of rhetorical device.

So, what Nozick claims is that there are a number of "non-experiential" things that are objectively important to us (e.g., to be importantly connected to reality), and that the experience machine is inadequate because it doesn't give us these things. But how should claims of this type be understood? For example, how should we interpret the idea that it is objectively important to us to be connected to reality? Well, let us first note that he does not necessarily mean that it has final value for us to be connected to reality, nor that it is always (not even *ceteris paribus*) nonderivatively better for us to be connected to reality than to be "disconnected" from reality. The reason why interpretations of this type are not open to us is that his claims are (it seems) consistent with modified hedonism, i.e. with Parfit's idea that "what is best for people is a composite" (cf. p 109 above). If we take this into account, we can understand his claim as follows: It has "contributory value" for a person to be connected to reality, i.e. a life cannot be good unless the person who is living this life is connected to reality, unless it (the life) is lived in actual connection with reality.

So, if we understand Nozick's argument in this way, is it a good

argument against pure hedonism? That is, can it be plausibly assumed that a person's well-being is (normally, in most cases) not only a function of how his life feels to him from the inside, e.g., that it is also directly dependent on whether he is connected to reality or not, and whether he lives his life autonomously or not? I think it can, but it is not easy not provide any strong arguments for this view (and as far as I can see, Nozick does not even try)[28].

A short note on Mill's Pig

It should not surprise us that Nozick's argument is consistent with modified hedonism. After all, his argument is (structurally speaking) almost identical with an argument that Mill (himself a modified hedonist) gives against pure hedonism, viz. the following one: Most of us would not consent "to be changed into any of the lower animals, for a promise of the fullest allowance of a beast's pleasures", or "to be a fool", or "an ignoramus", or "selfish and base", even if we would be "persuaded that the fool, the dunce, or the rascal is better satisfied with his lot" than we are with ours (Mill (1863), ch. II, p 8). In short, "[i]t is better to be a human being dissatisfied than a pig satisfied; better to be a Socrates dissatisfied than a fool satisfied" (ibid., p 9)[29].

Now, this argument must (obviously) be regarded as an argument against (H6), the idea that the value that a life has for the person who is living it is a function of one thing only, viz. how much pleasure and suffering it contains. What the argument shows us (if successful) is that a life which contains nothing but lower pleasures cannot be a good life (or more specifically, a good *human* life) for the person who is living it, or alternatively, that a human life is not a good human life unless it also contains higher pleasures, e.g., pleasures stemming from successful use of our "higher", characteristically human, capacities (cf. Tranøy (1973)). As far as I can see, this is really all that Mill purports (and manages) to show: "A life full of lower pleasures but devoid of higher pleasures is not good enough, and you should (for this reason) not always choose the lower of two pleasures". That is, he does not give us any reason to believe that a life full of higher pleasures but devoid of lower pleasures

---

[28]There has been attempts to justify different versions of the objective list theory, however. We will return to this issue in chapter 7.

[29]There is one difference between the two arguments, however: Both Socrates and the fool are (presumably) in contact with reality.

(assuming that such a life is possible) would be a good life, and he does not give us any reason to accept the view that Smart (1973) attributes to him, viz. when he writes that "Mill would wish to say that the pleasures of the philosopher were more valuable intrinsically than those of the dog, however intense these last might be" (p 17)[30].

Now, it is important to note that if Mill's argument is valid, it also hits (H5), the idea that the value of an experience for the person who has it is proportional to how pleasant or unpleasant it is. But it is far from obvious with what Mill would replace this claim, e.g., as I see it, it is far from certain that Mill's view implies that if two pleasures are equally pleasant, and if one is higher while the other is lower, then it is *always* (under all conditions) better for a person to have the higher pleasure. (That is, I don't think that Mill's view forces him to accept (c) on p 33). In any case, Mill's primary claim is a claim about the good *life*, and we should be careful when we try to draw conclusions about the value-for of particular experiences from such "holistic" claims.

HAPPY

Think of some "non-experiential" kind of situation that you strongly want (intrinsically) to be a "part of your life". It may be to be with a loved person, to have an intimate and reciprocal relationship, to engage in some creative activity, to act morally, to develop as a person, to achieve something important, or the like. For simplicity's sake, call this intrinsically desired situation X.

Now, imagine that you accept the hedonistic theory. You then believe that X can only be good for you in the instrumental sense (this is so, even if you happen to desire it intrinsically). X is good for you because it contributes to your happiness, e.g., because you take pleasure in it. Now, ask yourself: If X would no longer make you happy, would you stop regarding it as good for you? If the answer is "yes", this gives some support to the hedonistic theory, but this support is far from conclusive. You must also ask yourself whether it would make your life better if you would give up X for some small increase in happiness[31]. Let us assume that you are now reasonably happy, partly because X is the case. Now, imagine that you would be offered (for free) a harmless

---

[30]Personally, I don't think Mill held this view, i.e. I think he would accept (e) on pp 33-34 above.

[31]Cf. pro-argument (3) on p 110!

drug called HAPPY. If you take HAPPY, you will be happier than you are now, but you will also lose interest in X. X will not concern you anymore, you will simply stop wanting that X obtains, and this is also the reason why you will become happier. Would you be better off if you take HAPPY than if you don't?

The same story can be repeated for "non-experiential" things that you think are bad for you. This could be being totally deceived, being enslaved, being totally passive or utterly lonely. Call this bad thing Y. Here, taking HAPPY will change your attitude toward Y in a special way. It will make you start liking Y, and as a result of this, you will become happier than you would otherwise have been. In this case, would you be better off if you take HAPPY?

If you say no to HAPPY[32], you are not really a hedonist: the experiential dimension is not the only dimension you really care about. Most of us want more than to be happy, we also want to be happy "for the right reason" or "on the right grounds". We want, I think, certain kinds of wholes, where happiness is only one (let alone essential) component. And the reason why we want this is that we have realized that pure hedonism is an implausible theory.

We have now looked at all the arguments in the second group, and it is now time to turn to the third group of arguments, viz. the arguments that are specifically directed against (H3), the thesis of unrestrictedness, and/or (H4)-(H5).

Arguments against pure hedonism that are, at the same time, arguments for some modified version of the hedonistic theory: The criticism of pleasures

There are two kinds of arguments in this group. Arguments of the first kind are directed against (H3), the thesis of Unrestrictedness, i.e. they purport to show that (RH3) is plausible, that we should accept the idea that there are pleasant experiences that are not nonderivatively good for the experiencing subject, and that there might even be pleasures which are bad for the subject. (As far as I know, there has been no attempts to show that there are unpleasant experiences that are not

[32]*Qua* self-interested, that is; it is of course possible to say no to HAPPY for other (e.g., moral) reasons; just as it is possible to refuse to plug into the experience machine for moral reasons.

nonderivatively bad for the suffering subject). More specifically, all arguments of this kind can be viewed as attempts to establish that some *restriction claim* is plausible, i.e. that there is good reason to accept at least some claim of the form "If a pleasant experience has the 'non-hedonic' feature F (e.g., if it is irrational), then it is not nonderivatively good for the subject to have it, in spite of its pleasantness".

Arguments of the second kind are directed against (H4)-(H5), i.e. they purport to show that (MH4) is correct, that we should accept the idea that the nonderivative value-for of an experience is not just dependent on how pleasant or unpleasant this experience is, but also on what other features it has. More specifically, all arguments of this second kind can be viewed as attempts to demonstrate that there is good reason to accept some *claim concerning relative weights*. If we have pleasures in mind, such claims are of the following form: "If the pleasant experience E1 has the non-hedonic feature F1, and if the pleasant experience E2 has the non-hedonic feature F2, then it may be nonderivatively better for the subject to have E1, even if E2 is more pleasant than E1. And if E1 and E2 are equally pleasant, it is better to have E1 than to have E2. (And the reason for this is that E1 has F1 while E2 has F2)"[33].

Now, the number of possible restriction claims is considerable, and so is the number of possible claims about relative weights, and it would take up too much space to go through them all. So what I will do is this: I will restrict my attention to a few claims of each kind, viz. the claims that I myself find most intuitively plausible. I will then ask whether these claims are justified, i.e. whether there are good reasons for accepting them[34].

In what follows, I will mainly be concerned with pleasant *emotions*, i.e. with pleasures which have intentional objects, or alternatively, with

---

[33]Similar claims can (of course) be made for sufferings, and it is not impossible that some claims of this type are plausible. In this context, I will ignore the case of displeasure, however, in part because unpleasantness seems to have a more dominant standing among the bad-making features (of experiences) than pleasantness has among the good-making features.

[34]So, is it really sufficient for our purposes to restrict our attention to a few claims? Well, if the purpose is to refute pure hedonism, it is: after all, all we need in order to achieve this purpose is to establish that there is at least one valid restriction claim or claim concerning relative weights. But if the purpose is to find the best theory of prudential value, it might not be sufficient: Or more specifically, if some form of modified hedonism is the best theory, it is not sufficient to keep things this short.

pleasures which can be described as taking pleasure *in something*. In this context, there are a few things that need to be pointed out, viz.:

(i) The object that we take pleasure in when we take pleasure in something is not necessarily real. Roughly speaking, we do not take pleasure in how things actually are (e.g., in what is actually the case), but in how they appear to us (e.g., in what appears to be the case).

(ii) The objects of our pleasant emotions are often situations, or more precisely (but still roughly), what we believe is (or was, or will be) the case. These situations can be real or imagined, and they can be past, present, or future.

(iii) A type of "taking pleasure in" that is of particular interest in this context is "taking pleasure in doing something"[35]. It is therefore fruitful to make a distinction between those of a person's pleasant emotions that are "directed at" his own activity and those of his emotions which have other objects.

The restriction claims that are (on my view) most worth considering are:

> **(R1)** If a pleasant experience is associated with some morally unacceptable act or activity (on the part of the experiencing subject), then it is not nonderivatively good for this subject to have the experience; it may even be bad for him. Examples of such immoral activities (activities that it is, on this type of view, not good for a person to take pleasure in) are: The intentional infliction of pain or mental suffering on another sentient being, at least when this is done in order to feel pleasure or sexual gratification (and not, e.g., in order to get revenge), i.e. pure sadistic activity; actions which are manifestations of some moral vice (like greed); and actions that are motivated by pure malice[36].

---

[35]This is one way in which a pleasure can be associated with an activity; the other is that engaging in a certain activity can give rise to pleasure (or more precisely: that P's experience of engaging in an activity can cause that P feels pleasure).

[36]The idea that the value that a pleasant experience has for the experiencing subject is dependent on the *moral status* of the activity with which it is associated can be regarded as a version of a more general idea, viz. the Aristotelian idea that the nonderivative value-for of a pleasant experience is (at least in part) a function of the *value* of the activity to which it attaches (the activity which it "completes"). According to Annas (1980), this is "Aristotle's main thesis about pleasure and goodness: pleasures vary in goodness with the activities that give rise to them" (p 292). If we combine this idea with the idea that an activity is good if and only if it is excellent (or "virtuous"), i.e. if it is a manifestation of some excellent trait of

(R2) If the object of a pleasant emotion is an "objectively unpleasant" situation[37], then it is not nonderivatively good for the subject to have the emotion; it may even be bad for him to have it, in spite of its pleasantness. Examples of such "objectively unpleasant" situations (situations that might, on this type of view, not be worth taking pleasure in) are: The fact that someone else suffers, e.g., as a result of being tortured or humiliated by a third party; or the fact that the experiencing subject himself is humiliated, deceived, dominated, or degraded. That is, neither sadistic nor masochistic pleasure are (on this view) regarded as good for a person[38].

If someone doubts that it makes sense to say about a certain situation (e.g., an activity) that it is objectively unpleasant, i.e. that it is not worth taking pleasure in it, even though most people would (in fact) take pleasure in it, it might be helpful, as a comparison, to look at fear and danger: If a person P is afraid of something X, then X appears to P as dangerous, i.e. X is subjectively, or apparently, dangerous. It may not be objectively dangerous, however; it is a well-known fact that we are sometimes afraid of things that are not really dangerous, and that we are not always afraid of things that are really dangerous. That is, as far as the dangerous is concerned, the distinction between the objective and the subjective makes perfectly good sense. So why would it not make sense to distinguish between the subjectively pleasant and the objectively pleasant? On my view, it seems that just as there are things or situations which it is appropriate to be afraid of (i.e. objective dangers), there may be activities or situations which are "not worth taking pleasure in", or even "worth taking displeasure in", i.e. things

---

character (like courage, truthfulness, or justice), and that an activity is bad if and only if it is a manifestation of some fault (or flaw) of character (like cowardice, self-indulgence, or undue humility), we can at least "deduce" the following claim concerning relative weights: "If a pleasant experience E1 is associated with an excellent activity, and if another pleasant experience E2 corresponds to a flawed (or defect) activity, then E1 is *ceteris paribus* the better pleasure".

[37]The situations I have in mind here do not include the subject's own actions, i.e. facts of the type "the subject performs a certain action, or engages in a certain activity" are excluded.

[38]Is it as plausible to claim that the negative value of an unpleasant experience can be a function of its "moral status"? For example, is it nonderivatively worse for a person to be unhappy because someone else is happy than to be unhappy because someone else is unhappy?

which are objectively unpleasant.

So, are (R1) and (R2) justified claims, and if they are, how can they be justified? Notice that the question is *not* whether it is nonderivatively good-*period* that someone gets a kick out of engaging in sadistic activity, but whether it is not-good (e.g., bad) *for the sadist* to take pleasure in harming someone else, and if so, why. Assuming that there is such a thing as the objectively unpleasant, why should we attach any importance to it in this context? Harming another sentient being may well be objectively unpleasant, but the sadist does (after all) take pleasure in what he is doing, and isn't this what is of relevance here?

So, on what grounds can one reasonably claim that it is not good for a person to feel sadistic or masochistic pleasure? Well, apart from appealing directly to people's "local" intuitions, there are (as I see it) only two things we can do here, viz. (i) we can appeal to people's intuition about the good life (like the counter-arguments of the second type did), or (ii) we can appeal to some conception of the well-functioning person, i.e. we can claim that sadists and masochists are "abnormal" ("sick", or "crazy") people, and that their pleasures are "abnormal" pleasures.

The first claim concerning relative weights that is (on my view) worth considering is this:

> **(RW1)** If the pleasant emotion E1 is appropriate, and if the pleasant emotion E2 is inappropriate, then it may be nonderivatively better for the subject to have E1, even if E2 is more pleasant than E1. And if two emotions are equally pleasant, then it is better to have the emotion that is more appropriate.

So, what does the appropriateness (inappropriateness) of an emotion consist in? If we accept Nozick's (1989) idea that an emotion consists of a belief, an evaluation, and a bodily feeling (cf. note 12 on p 77 above), we can say (still following Nozick) that

> [a]n emotion can be defective or inappropriate in three ways: the belief can be false; the evaluation can be false or wrong; or the feeling can be disproportionate to the evaluation. /.../
>
> Let us say that an emotion fits when it has the /.../ threefold structure of belief, evaluation, and feeling, and moreover when the belief is true, the evaluation is correct [informed, unbiased, supported

by reasons, justified, or whatever], and the feeling is proportionate to the evaluation (p 89).

On this view, the notions of appropriateness and inappropriateness are both "three-dimensional". An emotion is appropriate in the "belief-dimension" if its constituent beliefs are true, and inappropriate if they are false; it is appropriate in the "evaluative dimension" if its constituent evaluation is correct (justified, etc.), and inappropriate if it is wrong (etc.); and it is appropriate in the third dimension if its constituent feeling is "proportionate to its constituent evaluation", and inappropriate if the feeling is "disproportionate to the evaluation".

This suggests that (RW1) can be broken down into the three claims (assuming that all three dimensions are of relevance, that is); (RW2), (RW3), and (RW4).

(RW2) If the pleasant emotion E1 is based on a true belief, and if the pleasant emotion E2 is based on a false belief, then it may be intrinsically better for the subject to have E1, even if E2 is more pleasant than E1 (and so on). For example, it is *ceteris paribus* better for a person to take pleasure in what is actually the case than to take pleasure in something that is not the case.

(RW3) If the pleasant emotion E1 is based on a correct (justified) evaluation, and if the pleasant emotion E2 is based on an incorrect (unjustified) evaluation, then it may be nonderivatively better for the subject to have E1, even if E2 is more pleasant than E1. Or alternatively (if we assume that the evaluation "X is good in way W" is correct if and only if X is (in fact) good in this way): It is *ceteris paribus* better for a person to take pleasure in a valuable object than to take pleasure in something that is not valuable[39].

---

[39]Note that this claim is not identical with the claim that if a pleasant emotion E1 has a *more valuable object* than another pleasant emotion E2, then it may be nonderivatively better for the subject to have E1, even if E2 is more pleasant than E1. Now, this idea is a version of the more general idea that the value of a pleasant intentional mental state (e.g., an emotion, a perception, a fantasy, or a thought) is a function of how valuable its object is. This idea can also be formulated in Aristotelian terms, i.e. in terms of activity. If we do this, we get: "If a pleasant experience is associated with an intentional mental activity (e.g., like emoting, perceiving, fantasizing, or thinking), then the value of this experience is (in part) a function of the (objective) value of the object of the activity". This view was held by Aristotle himself, who believed (i) that the value of a pleasure is a function of the value of the activity with which it is associated, and (ii) that if the

As far as I can see, the constituent evaluation of an emotion can be of many kinds, e.g., it can be of the form "X (the object) is beautiful", or of the form "X is nonderivatively or derivatively good for me (him, her)", or of the form "X has final or instrumental value-period". That is, the idea is that if a person takes pleasure in something because it appears to him as beautiful, it is *ceteris paribus* better for him if the object of his pleasant emotion is in fact beautiful rather than ugly (assuming that beauty and ugliness are *not* in the eye of the beholder), and if a person takes pleasure in something because it appears to him as good-for-him, it is *ceteris paribus* better for him if the object of his pleasant emotion is in fact good-for-him rather than bad-for-him.

(RW4) If E1 is a pleasant emotion whose constituent feeling is more proportionate to its constituent evaluation, and if E2 is a pleasant emotion whose constituent feeling is less proportionate to its constituent evaluation, then it may be nonderivatively better for the subject to have E1, even if E2 is more pleasant than E1. And if two emotions are equally pleasant, then it is *ceteris paribus* better to have the emotion whose constituent feeling is more proportionate to its consti-

---

activity with which a pleasure is associated is a mental activity with an intentional object, e.g. an activity of one of the different senses, then the value of the activity is a function of two things, viz. the value of the object and the value of the "organ". This is Aristotle's own formulation of (ii) (in *NE*, book X, chapter 4): "Since every sense is active in relation to its object, and a sense which is in good condition acts perfectly in relation to the most beautiful of its objects /.../, it follows that in the case of each sense the best activity is that of the best-conditioned organ in relation to the finest of its objects. And this activity will be the most complete and pleasant. For, while there is pleasure in respect of any sense, and in respect of thought and contemplation no less, the most complete is pleasant, and that of a well-conditioned organ in relation to the worthiest of its objects is the most complete; and the pleasure completes the activity (p 256). For example, if a person takes pleasure in thinking about something, the value of this pleasure is a function of (a) how good a thinker he is, and (b) how valuable the object of his thought is, e.g., it is (on Aristotle's view) better to take pleasure in thinking about eternal things than in thinking about human affairs: "there are other things much more divine in their nature even than man, e.g. most conspicuously, the bodies of which the heavens are framed" (*NE*, book VI, chapter 7). And if someone takes pleasure in looking at something, the value of this pleasure is a function of (a) how well he can see (how "well-conditioned" his "eyes" are), and (b) how beautiful the object is. That is, if "the eye of the beholder" remains the same, it is better for him to take pleasure in looking at something objectively beautiful than to take pleasure in looking at something objectively ugly.

tuent evaluation.

So, what is it for a feeling to be proportionate (disproportionate) to an evaluation? This is an example, given by Nozick (1989), of disproportion:

> Suppose, walking along the street, I find a dollar bill and feel ecstatic. You ask whether I think it indicates this is my lucky day or that my fortunes have changed or that I am beloved by the gods, but no, it is none of these things. I simply am ecstatic. But finding a dollar bill isn't *that* wonderful a thing; the strength and the intensity of the feeling should bear some proportionate relationship to the evaluation of how good a thing finding a dollar is - to the measure of the evaluation (p 89)[40].

Here, it is important to keep in mind that there is also another way in which a feeling can be disproportionate to an evaluation, viz. that it can also be "too weak" in relation to the evaluation. An example of this kind of disproportion is when someone (e.g., because of depression or some deep-seated inhibition) is unable to take pleasure in the fact that his love is reciprocated, or in the fact that he got the job he values the most[41].

The last claim concerning relative weights that we will consider is this:

> **(RW5)** If the pleasant emotion E1 is directed at a real (actual, present) object, and if the pleasant emotion E2 has an imaginary (unreal) object (e.g., if E2 is an anticipatory pleasure), then it may be nonderivatively better for the subject to have E1, even if E2 is more pleasant than E1.

In order to clarify this idea, let us distinguish between the following three types of cases:

(a) A person takes pleasure in something that is actually the case (right now), e.g., he is satisfied with his job, or he enjoys his holiday fully.

---

[40]Where the term "I" seems to refer to a normal adult, and not to a small child.

[41]A similar claim could be made for sufferings, viz. that it is *ceteris paribus* less bad to have an unpleasant emotion whose constituent feeling is more proportionate to its constituent evaluation. For example, it might be claimed that the reason why it is bad for a person to suffer intensely from an ordinary mosquito-bite is not just that it is unpleasant, but also that the feeling is disproportionate, or it might be claimed that it is not always better for a person to grieve the loss of a loved one less, e.g., that it may even be bad for a person if he (because of some deep-seated inhibition) is unable to grieve the loss of a loved one.

(b) A person takes pleasure in something that is no longer the case, e.g., he remembers with pleasure how much fun he had when he was a teenager, or when he was in Greece last year.

(c) A person takes pleasure in an imaginary state, e.g., by day-dreaming about a perfect relationship or about next Summer (when the weather will be perfect and there will be no insects around). Many of the pleasures which belong to this category are "anticipatory pleasures", i.e. the imaginary situations in which we take pleasures are often "future situations".

Now, the idea is that it is *ceteris paribus* better for a person to have pleasures of type (a) than to have pleasures of types (b) and (c)[42]. (That is, there is an "extensional overlap" between (RW5) and (RW2), but the two are not identical).

So, are (RW2)-(RW5) justified claims? The pure hedonist thinks not, and it is likely that he bases his rejection of the claims on the following type of argument: "It is true that it is often better for a person to have pleasures that are based on true (rather than false) beliefs, or to take pleasures in real (rather than imaginary) things, but it is hardly non-derivatively better for him. The belief that it is is based on a confusion; if we think clearly, we will see that the reason why we regard certain pleasures as superior to others is really that they are more permanent, stable, fecund, or the like"[43].

So, what arguments can be given for (RW2)-(RW5), e.g., what reasons are there for believing that it is *ceteris paribus* nonderivatively better for a person to have pleasant emotions whose component beliefs are true than to have pleasant emotions whose component beliefs are false (and so on)? Well, I don't think it is really possible to come up with any arguments that the pure hedonist would accept. What we can do is to appeal to different kinds of intuitions, e.g., we can construct concrete examples that appeal to our intuitions about particular cases, we can construct arguments that appeal to our intuitions about good lives, or we can appeal to our intuitions about the well-functioning person (cf. p 146 above).

---

[42]It is worth noting that there are objects of pleasant emotions which are hard to classify as "real" or "imaginary". For example, if someone take pleasure in the fact that he has certain abilities and/or opportunities, to which category does this pleasure belong?

[43]Cf. note 3 on p 108 above.

For example, this is how one might argue for (RW2), the idea that it is *ceteris paribus* nonderivatively better for a person if the belief component of a pleasant emotion is true than if it is false: Consider (again) Rachel's story about Wonmug. Wonmug took pleasure in achieving great things, but he really achieved nothing, and he took pleasure in having many friends, but he didn't really have any friends. That is, Wonmug's pleasures were (primarily in a conceptual sense, but also in a causal sense) based on a number of false beliefs, e.g., the false belief that he was (in fact) achieving great things, and the false belief that he had many good friends. Now, Wonmug never discovered that his beliefs were false, but it would (on my view) still have been better for Wonmug if his beliefs were true, i.e. if he had really achieved something and if he really had friends. But why would this be better for him? Well, in part because it would have been good for him to achieve great things and have good friends (either because he desired to have these things, or because achievement and friendship are objective prudential values)[44], but this is not the whole explanation: It is also good for a person to be in contact with reality, e.g., to have true beliefs[45], and to have such contact is a constituent "part" of being a well-functioning person.

This is how one might argue for (RW4), the idea that it is *ceteris paribus* better for a person to have a pleasant emotion whose constituent feeling is proportionate to its constituent evaluation than to have a pleasant emotion whose constituent feeling is disproportionate to its constituent evaluation: Let us stick to Nozick's example, and ask: Would it not be better for the person if he felt just as ecstatic about

---

[44]That is, it is not all about having true beliefs, e.g., it is also necessary to take (RW3) into account. To see this, consider a person who takes pleasure in something bad, like being humiliated, rather than in something good, like having many good friends. Now, compare the following possible states: In State 1, the person is happy because he believes he has just been humiliated, but his belief is false, and in State 2, he is happy for the same reason, but this time, his belief is correct, i.e. he is happy because he *knows* that he has just been humiliated. In this case, it might well be better for the person if State 1 holds, i.e. if he has, in fact, not been humiliated. But this does not show that it does not matter whether his beliefs are true or not; all it shows is that the prudential goodness of having true beliefs can sometimes be outweighed by other objective prudential values, e.g., by the fact that it is bad for a person to be humiliated.

[45]That is, there are two reasons why it might be good for P that his belief that he has friends is true: (i) it is good for him to have true beliefs, and (ii) it is good for him to have friends (Where (ii) is probably more important from P's own (immanent) perspective).

something more valuable than finding a dollar, e.g., getting the job he had applied for? Of course it would, but this intuition is better explained by (RW3) than by (RW4), i.e. we seem to need a better test. So, let us instead ask: Would it be better for the person who found the dollar if he would have been less ecstatic? I think it would, and this belief is (again) based on the idea that it is (in general) better for a person to be in contact with reality than to be "out of touch" with reality.

My argument for (RW5) also appeals to the idea that it is (in most cases) good for a person to be in contact with reality. But notice that (RW5) does not follow from this idea in any straightforward manner; to take pleasure in something that is no longer the case is, after all, not to be "out of touch" with reality, and the same thing holds for anticipatory pleasure; anticipating future events with pleasure is hardly incompatible with being in contact with reality. However, what we can say is that a person who tends to dwell in the past or the future much more than in the present is not really in touch with reality, and if he often takes pleasure in what is no longer that case, or in imaginary future situations, but never in what is actually the case, then this circumstance detracts from the value of his life.

This ends the section on the arguments against (pure) hedonism.

Conclusion

To conclude, there are a number of strong arguments against pure hedonism, and this theory should therefore be rejected. The strongest of these arguments do not hit modified hedonism, however, and we should therefore be reluctant to reject this theory. In fact, it seems that this is a theory of prudential value that can easily incorporate two very plausible claims, viz. (i) the idea that a theory of well-being (the good life) can not be plausible unless it includes a number of non-hedonistic components, and (ii) the idea that a complex situation (a whole) cannot have nonderivative value for a person unless it has some pleasant experiential content. One possible version of this "modified hedonistic position" has been formulated by Parfit (1984)[46]:

> What is good for someone is neither just what Hedonist's claim, nor just what is claimed by Objective List Theorists. We might believe that

[46]Cf. also p 109 above.

> if we had *either* of these, *without the other*, what we had would have little or no value. We might claim, for example, that what is good or bad for someone is to have knowledge, to be engaged in rational activity, to experience mutual love, and to be aware of beauty, while [taking pleasure in] /.../ these things. On this view, each side in this disagreement saw only half of the truth. Each put forward as sufficient something that was only necessary. Pleasure with many other kinds of object has no value. And, if they are entirely devoid of pleasure, there is no value in knowledge, rational activity, love, or the awareness of beauty. What is of value, or is good for someone, is to have both /.../ (p 502).

This may well be a plausible theory of prudential value, but in order to find out whether this is really the case (i.e. in order to find out whether (ii) is a plausible claim), we first have to take a closer look at the remaining two theories, viz. the desire-fulfilment theory and the objective list theory. We will begin with the former.

Chapter Four
A formulation of the Unrestricted Desire-Fulfilment Theory

The discussion of the desire theory will have the following structure. In the present chapter, I will try to give a formulation of the unrestricted desire theory that is as precise as possible[1]. In connection with this, several topics will be discussed, e.g., what it is for someone to desire something, what it is for a desire to be stronger than another desire, what it is to have a desire fulfilled (or satisfied), and what it is for a desire to be intrinsic. The central questions in chapter 5 are questions of plausibility. By looking at a number of arguments that can be given for and against different versions of the desire theory, I will try to find out (i) what version of the theory that is the most plausible theory of prudential value, and (ii) whether any version of the theory is plausible. The chapter will start with a critical discussion of the unrestricted desire theory, but other (modified) versions of the desire theory will be introduced as we move along. The reason why I choose this "order of presentation" (why I do not formulate all the versions first, and discuss them later) is that the other versions of the theory are best understood as modifications of the unrestricted theory; modifications that have been made in order to deal with certain objections that have been directed against the unrestricted theory.

To recapitulate how the satisfaction interpretation of the unrestricted desire theory was characterized in section 1.2, here is a brief summary:

(D1) Nothing but (actual) desire-fulfilment can be nonderivatively good for a person, and nothing but "aversion-fulfilment" can be nonderivatively bad for a person.

---

[1]The reason why I will start from this version of the desire theory is that it is, in a certain sense, the most fundamental version of the theory; see below.

(UD2) The thesis of Unrestrictedness: There are no (intrinsic) desires that it is not nonderivatively good for a person to have fulfilled, and there are no (intrinsic) aversions that it is not nonderivatively bad to have "fulfilled".

(UD3) To the extent that it is possible to determine just how strong our desires and aversions are: The positive (or negative) value that a certain desire-fulfilment (or aversion-fulfilment) has for a certain desiring subject is proportional to how strong the desire (or aversion) is. That is, to the extent that desires and aversion are comparable with respect to strength: The stronger (the more "intense") an intrinsic desire (aversion) is, the nonderivatively better (worse) it is for the desiring subject to have it fulfilled. This is the "intensity-orientation" of the theory.

(D1)-(UD3) are claims about the value-for of local situations (or "atomic facts"). (UD4), on the other hand, is a claim about the value-for of whole lives-at-certain-times, situations the value-for of which cannot be determined directly.

(UD4) To the extent that it is possible to determine how well off a certain person is: The value that a certain life has for the person who is living it is a function of how much desire-fulfilment and how much aversion-fulfilment this life "contains". The more desire-fulfilment and the less aversion-fulfilment a life contains, the better this life is for the person who lives it[2].

As it stands, this claim is far too imprecise. So, how would an unrestricted desire theorist make it more precise? Or alternatively put, how exactly should we determine how well off someone is: how should the function in question be characterized? Here, it is likely that our unrestricted desire theorist would use the same approach as the pure hedonist (see pp 69-71), i.e. that he would adopt the following position:

(UD4:a) First, he would (as we have seen) assume that "evaluative atomism" is true, i.e. that the final value-for of a person's existence is a function of the nonderivative values-for (positive and negative) of

[2]Here, it might be tempting to say that the value that P's life has for P is a function of to what extent (to what degree) his preferences are satisfied, but as we will soon see, this would be a big mistake.

its constituent parts.

So, how do we calculate a person's level of well-being at a certain time from the positive and negative values-for of the particular desire-fulfilments and aversion-fulfilments that his life "contains" at that time? In short, how does the desire theorist tackle the problem of aggregation? Well, as a first approximation, we know that the unrestricted desire theorist would adopt the following rudimentary theory of aggregation: The more desire-fulfilment and the less aversion-fulfilment a certain life contains, the better for the person who is living it. For example, a person's well-being is improved if there is an increase in the "amount of desire-fulfilment" that his life contains, while the "amount of aversion-fulfilment" remains the same.

But how do we determine how much desire-fulfilment or aversion-fulfilment a life contains? And how do we deal with the cases where one possible life contains both more desire-fulfilment and more aversion-fulfilment than another possible life? The desire theoretical answer to these questions can (just like the hedonistic answer) be divided into two parts:

(UD4:b) This is the procedure that should (in principle, and if possible) be adopted in order to determine how much desire-fulfilment and how much aversion-fulfilment a certain life contains: We assign a positive number to each desire-fulfilment, and a negative number to each aversion-fulfilment, that the life contains[3]. (How great these numbers are depends on how strong the desires and aversions are). We then add all the positive numbers we have assigned to the desire-fulfilments, and all the negative numbers we have assigned to the aversion-fulfilments. If we do this, we get two sums, viz. the positive sum Σ (df), and the negative sum Σ (af). We can now say that the higher the sum Σ (df) is, the more desire-fulfilment a life contains, and the lower (the more negative) the negative sum Σ (af) is, the higher the amount of aversion-fulfilment[4].

---

[3]Assuming that a person's desires and aversions can be individuated in a satisfactory way, that is. We will return to "the problem of individuation" on pp 158-159 below.

[4]It is important to point out that for this type of aggregation to be at all possible, two assumptions have to be made, viz. (i) that the strengths of our desires and aversions are measurable on ratio scales, and (ii) (which is more dubious) that the values-for of our desire-fulfilments and aversion-fulfilments satisfy "the criterion

Here, it is worth pointing out that we can safely ignore how many desires and aversions that are *not* fulfilled. The reason for this is simple: The idea that we cannot calculate the value-for of a person P's life unless we take into account how many desires that are not fulfilled (and how strong these desires are) is based on the idea that the value that P's life has for P is a function of *to what extent* his desires are satisfied. This idea is mistaken, however, and the reason for this is that it is based on the dubious assumption that it is bad for a person not to have his desires fulfilled (and good for a person not to have his aversions fulfilled). In short, when we realize that non-fulfilment (of desires and aversions) is neither good nor bad for a person, we also realize that it does not matter *per se* to what degree his desires and aversions are satisfied. That is, the only reason why it is better for a person to have all his desires fulfilled than to have only some of them fulfilled is that the number of desires fulfilled is higher in the former case, and it is not always better for a person to have few desires and have them all fulfilled than to have many desires and only have some of them fulfilled[5] [6].

(UD4:c) Once our desire theorist has access to the two sums Σ (df) and Σ (af), he can make the idea that "the more desire-fulfilment and the less aversion-fulfilment that a certain life contains, the better for the person who is living this life" more precise, viz. as follows: The

---

of additivity" (i.e. that the numerical values assigned can be meaningfully added).

[5]This is in line with what Nordenfelt (1991) says, but he expresses himself in an unfortunate way. If we focus on the desire-theoretical component of his theory of "happiness", we find the following view: "Happiness" has two dimensions, and how happy a person is (on the whole) is a function of how happy he is in these dimensions. The first dimension is the "equilibrium dimension", and how happy a person is in this dimension depends on to what degree his desires are fulfilled (on how large a proportion of his desires that are fulfilled). The second dimension is "the richness dimension", and how happy a person is in this dimension depends (roughly) on how much he desires (on how many and how "ambitious" his desires are). But instead of complicating the issue in this way, why not say right away that it is the "amount of desire-fulfilment" that counts?! (It is possible that I am a bit unfair here; it is likely that Nordenfelt has accepted a phenomenological conception of desire, a conception on which it might well be bad for a person not to have his desires fulfilled. This could also explain why he does not treat our aversions separately).

[6]All this has an interesting practical implication: If we accept some "non-phenomenological version" of the desire theory, and if we want to improve our lives, we should not use the well-known strategy of trying to minimize our desires; what we should try to minimize is our aversions!

greater Σ (df) is, the better, and the "smaller" (the more negative) Σ (af) is, the worse. But how exactly should we (according to the desire theorist) calculate the nonderivative value-for of a life from its Σ (df) and Σ (af)?

There are at least two different ways in which an unrestricted desire theorist might answer this question, viz. he can either appeal to (1) the difference thesis, or to (2) some kind of ratio thesis.

(1) The Difference Thesis is the idea that a life L1 is nonderivatively better than another life L2 iff the difference [Σ (df1) - | Σ (af1) |] is bigger than the difference [Σ (df2) - | Σ (af2) |].

(2) This is a possible Ratio Thesis: A life L1 is better than another life L2 if and only if the ratio Σ (df1)/[Σ (df1) + | Σ (af1) |] is bigger than the ratio Σ (df2)/[Σ (df2) + | Σ (af2) |][7] [8].

It is worth noting that this formulation of (UD4) is open to a number of methodological objections, e.g., it might be claimed that at least some of the methodological assumptions on which it is based[9] are false. Now, even though all methodological issues have been more or less excluded from this book, there is one methodological difficulty that is actualized by (UD4) which is worth pointing out in this context, viz. "the problem of individuation": The idea that a person's overall level of well-being is a function of how many of his desires and aversions that are fulfilled (and how strong they are) can only make sense if it is (in principle)

---

[7]Again, it is important to notice that these "operations" are not meaningful unless the following two conditions are (in addition to (i) and (ii) in note 4) satisfied: (iii) "Desire-strength" and "aversion-strength" are measurable on *the same* ratio scale, i.e. the strengths of our desires are fully comparable to the strengths of our aversions, and (iv) it is not just that the two sums Σ (df) and Σ (af) can be "defined", it is also meaningful to perform arithmetic operations (like addition or division) on these sums (once they have been defined).

[8]If we restrict ourselves to desires and aversions that are roughly equal in strength, and if we assume that a person's desire-fulfilments and aversion-fulfilments can be counted (something that presupposes that our desires and aversions can be properly individuated), and if we let N(df) (and N(af)) denote the number of desires (aversions) fulfilled, then the difference thesis would claim that the bigger the difference between [N(df) - N(af)] is, the better, and the ratio thesis would claim that the bigger the ratio [N(df)/[N(df) + N(af)] is, the better. (That is, both theories imply that it is better if N(df) > N(af) than if N(df) = N(af); and that it is better if N(df) = N(af) than if N(df) < N(af)).

[9]In particular, (i)-(iv) in notes 4 and 7.

possible to separate a person's desires and aversions from each other[10]. This might be very difficult, however (it might even be impossible). This is one way in which the problem can be formulated: A complete description of a person's total desire-state at a certain time has to include a complete description of what it is that the person desires at that time. But it is not unlikely that such an "object-description" would contain certain overlaps, e.g., a certain person might (to take a simple example) want a car, but he might "also" want a Swedish car and a Volvo. Now, if we regard these desires as different, this implies that the same particular occurrence (he gets a Volvo) will constitute a fulfilment of several different desires at the same time. And if we (on top of this) accept (UD4), we would have to accept that it is better for me to get a Volvo if I want a Volvo and a Swedish car than if I just want a Volvo. But this is (as I see it) hardly plausible. This suggests that for (UD4) to be plausible, we have to require that the descriptions of the objects of a person's desires are somehow "independent", so that the same particular situation cannot be counted several times. So the question arises: Is it possible to give such a description of what a person desires? Does our ordinary language (on which our common sense individuation of our desires and aversions, i.e. in terms of propositional content, is based) have "what it takes"?[11]

Let us now turn our attention to (D1)-(UD3). As these claims stand, they are not sufficiently clear and precise for our purposes; in order to find out how plausible the different versions of the desire theory are, they have to be formulated in a more precise way. So, what is it that is in need of clarification here? What is it that we need to know in order to arrive at a sufficiently precise formulation of the unrestricted desire

---

[10]Assuming that it makes sense to think of a person's "total desire-state" at a certain time as consisting of a number of individual desires and aversions. We have already seen (on pp 79-81), that there is a similar problem in connection with hedonism, but it seems that the present problem is of a more serious kind. *If* it is correct to think of a desiring subject as having one big desire-state (a state that can be satisfied to different degrees), then is the most plausible way to analyze (or describe) this state really in terms of individual desires? Well, it seems that as long as we stick to ordinary language, there is no other way (cf. below).

[11]Another possible solution is this: "If a particular, concrete situation constitutes the fulfilment of several desires, count only (for purposes of aggregation) the strongest one". The problem with this suggestion is that it seems inconsistent with the fact that it is sometimes possible to kill two birds with one stone.

theory? On my view, there are at least four "sources of unclarity", viz. the following ones:

(1) When the unrestricted desire theorist claims that it is always non-derivatively good for us to have our (intrinsic) desires fulfilled, and bad for us to have our aversions fulfilled, it is not entirely clear how the terms "desire" and "aversion" should be understood. How should these key terms be interpreted in this context? What we want to know here is not just how the terms are actually used by the desire theorist, but also (and this is really the primary thing) how he should use them. That is, what we primarily want to know is which of the possible interpretations (or senses) of the terms "desire" and "aversion" that makes the desire theory most plausible, but with the following "proviso"; this interpretation must not deviate too much (especially not extensionally) from the ordinary use (or uses) of the terms, and it must be consistent with at least some of our common sense beliefs about desire, e.g., with the belief that we are sometimes mistaken about what we want. (This is mainly to secure that the resulting "theory" is a theory about *desire*, and not about something else).

These questions need not be formulated in semantic terms, however; they can also be formulated in terms of conceptions of desire (and aversion). A *conception of desire* is a theory that purports to give answers to certain central questions about desire, e.g., questions like: What is it for someone to desire something? Do all mental states which are properly conceived of as desires have anything in common, some property (intrinsic or relational) in virtue of which they are classified as desires, or is the class of desire a "heterogeneous class"? That is, are all desires of the same *genus* (in the traditional sense of the term) or not? What fundamental forms (e.g., *species*) of desire are there? And how are our desires connected to things like our intentions, motives, reasons, values, emotions and actions?

If we formulate our questions in terms of conceptions, we get: What conception of desire is it likely that the desire theorist has in mind? And (above all) which possible conception of desire makes (if combined with the desire theory) the theory most plausible?[12]

---

[12]It is also worth mentioning what we are *not* really interested in in this context. We are not looking for the true meaning (nominal definition) of the term "desire", a meaning which captures the real nature (or essence) of desire. Even if there were such a thing as the nature of desire (which is doubtful), and even if

(2) When the unrestricted desire theorist claims (roughly) that it is nonderivatively better for a person to have a stronger (intrinsic) desire fulfilled than to have a weaker desire fulfilled, or that the positive value that it has for a person to have a certain desire fulfilled is a function of how strong the desire is, it is not entirely clear how the terms "stronger" and "strong" should be understood. So, in order to understand these claims, we need to know how the term "strength" is actually used in this context. But above all, we want to know what notion of strength that makes an "intensity-oriented" desire theory most plausible[13].

(3) It has already been suggested (in section 1.2) that terms like "fulfilled" and "fulfilment" are normally understood in the following way: A person's desire that a situation X obtains is fulfilled if and only if he desires that X *and* X holds, and a person's aversion to a situation Y is fulfilled if and only if he has an aversion to Y *and* Y obtains[14]. This idea can also be expressed as follows: A desire or aversion is fulfilled if and only if its propositional content is true (this notion of fulfilment "follows" naturally from the idea that desires and aversions are propositional attitudes; cf. section 4.1 below).

Now, it is important to note that this "traditional" notion of fulfilment has at least two interesting implications, viz. the following ones: (i) A person can have a desire fulfilled without knowing that it is fulfilled. As Brandt (1979) points out, that a person's desires are satisfied means that the desired "states of affairs" occur, "with no implications about his enjoyment of, or even knowledge about, [these occurrences]" (pp 146-147). (ii) A desire can (it seems) be fulfilled after it is long gone. If a

---

knowledge about this "thing" would give us "the true meaning of the desire theory", we could still ignore it. What we want to know is not what desire *really is*, but what conception of desire that is most relevant in this particular context. Neither are we *particularly* interested in how the term "desire" is used in ordinary speech, nor in what conception of desire that is most useful for psychological theory.

[13]These questions are (of course) intimately connected to the questions in group (1), i.e. the answers given to (1) will put restrictions on what answers that can be given to (2), and vice versa.

[14]As far as I can see, this is not the only possible "definition" of what it is to have a desire fulfilled, e.g., one might also specify the notion of fulfilment as follows: "If P desires that X, then this desire is fulfilled if and only if X actually obtains". Notice that this formulation is (so to speak) weaker than the "explication" we have chosen, viz. "A person's desire that X is fulfilled if and only if he desires that X and X holds". The latter proposition implies the former, but not vice versa.

person desires at t_1 that X holds at some future time t_2, and if X in fact holds at t_2, then the desire he had at t_1 is fulfilled, even if he has lost the desire in the meantime. As Brandt writes: "Suppose Mr. *X* at a time *t* wants an occurrence *O* at some time *t'*, or at any of many moments t_i to t_n. Then, if *O* actually occurs at some one of these times, *X*'s desire has been satisfied" (ibid., p 249)[15].

This gives rise to the following questions: Is it really plausible to allow for the possibility that a person's well-being can be directly affected by things he does not know anything about (cf. pp 126-128 above)? And is it plausible to assume that it can be nonderivatively good for us to have our "now-for-then-desires" fulfilled? Or alternatively put, is the traditional notion of fulfilment really the notion which has most "moral and rational significance", or is there some alternative notion of fulfilment which makes the desire theory more plausible? For example, should the desire theorist replace the traditional notion of fulfilment with (i) the idea that P's desire that X is *fulfilled* (in the relevant sense) if and only if P desires that X, X holds, *and* the desire and its object are simultaneous (the time of the desire coincides with the time of the occurrence of X); or with (ii) the idea that P's desire that X is *fulfilled* (in the relevant sense) if and only if P desires that X, X holds, *and* P is aware of the occurrence of X; or with (iii) the combination of (i) and (ii)[16]?

In this chapter, I will restrict my attention to the following question: Which notion of fulfilment is more relevant, the traditional notion or (i)? Or alternatively put, is it really plausible to allow for the possibility that it is nonderivatively good for a person to have a prospective (or retrospective) desire fulfilled? (In connection with this, there will also be some discussion of some other "temporal issues" that are actualized by the desire theory)[17].

---

[15]There is also a third implication (which is of less importance in this context), viz. that desire-fulfilment is an all-or-nothing matter. That is, there is (on this view) no such thing as *degrees* of desire-fulfilment.

[16]This idea can be spelled out as follows: P's desire that X is *fulfilled (in the relevant sense)* if and only if the following four conditions are met: (a) P desires that X, (b) X holds, (c) the desire and X are simultaneous, and (d) P is aware of the occurrence of X.

[17]The alternative notion (ii) will not be discussed in this chapter, but in chapter 5. My reason for postponing this discussion is simple: This notion of fulfilment deviates too much from the notion which is actually adopted by desire theorists (it also deviates too much from the ordinary usage of the term "fulfilment), and it is (for this reason) doubtful whether a desire theory which incorporates (ii) can be

(4) The unrestricted desire theorist seems to assume that it can not be nonderivatively good for a person to have a desire fulfilled unless this desire is *intrinsic*. So, what is it for a desire to be intrinsic; what is it to desire something in an intrinsic way? And if there are several possible answers to this question: What notion of intrinsicality makes (if adopted) the "intrinsicality condition" most plausible?

A central problem that is intimately connected to (4) (but also to (2)) is this: We know that the positive nonderivative value that it has for a person to have an intrinsic desire fulfilled is (on the unrestricted desire theory) a function of how strong the desire is. But what about all those desires that are in part intrinsic and in part extrinsic (e.g., instrumental); how do we determine (in principle, that is) how nonderivatively good it is for a person to have *such* a desire fulfilled? Suppose a person P's desire that X and his desire that Y are both "mixed" in this way, i.e. suppose P has two kinds of reasons for wanting X and Y to obtain, viz. because these situations have the intrinsic properties they have (or the like), but also because he believes that they will have certain effects that are (on his view) desirable, or because he believes that they will make something desirable possible. In this type of case, how do we determine which of the two desires that it is nonderivatively better for P to have fulfilled? Well, supposedly by isolating "the intrinsic component" of the desires, and then determine which is stronger (in the relevant sense). But is it possible to separate the intrinsic component of a "mixed" desire (or preference) in this way, and if it is, how should it be done? For example, should we first ask P to imagine two possible worlds that are identical in all respects, apart from the fact that X (but not Y) obtains in the one world while Y (but not X) obtains in the other, and then ask him which of the two worlds he would prefer? Or should we try to specify the propositional content of the two desires in such a way (individuate the desires in such a way) so that we will end up with a larger number of "pure desires", all of which are either intrinsic or extrinsic? Personally, I tend to adopt the latter solution, but this is only a tentative view. In any case, the important thing here is to notice that the type of desire theory which has been formulated here presupposes that this problem can be solved, i.e. that it is (in principle) possible to isolate the intrinsic component of our mixed desires and preferences.

---

regarded as an unrestricted theory.

These are some important questions we have to answer in order to arrive at a precise enough formulation of the unrestricted desire theory. Let us start with the first one, i.e. the question of what a desire is.

4.1. What is a desire?

To begin with, desires are not separate entities; every desire is someone's desire. Moreover, to desire is always to desire something. In other words, when we think of desires (or wants, which will be regarded as the same thing), we must think of them as "embedded" in the following structure: "Someone (a subject) desires (or wants) something (an object)"[18]. In order to understand what a desire is, we need to take a closer look at the three elements of this structure, viz. the desiring subject, the object of desire, and the desiring "itself".

The desiring subject

In the present context, we are only interested in the desires of human beings. So, what can we say about the human subject of desire? Well, all that needs to be said here is that when we attribute desires to a human being, we always attribute desires to him or her *qua* conscious agent (or *qua* person, in some minimal sense). To be able to attribute full-fledged desires to a human being or some other animal, it is necessary that we view him or her from "the psychological perspective", i.e. (roughly) as a conscious agent, where the term "agent" should be understood as "capable of voluntary action" (or something of the sort), and not merely as "intentional system", or "someone or something that displays goal-oriented behaviour"[19].

This is not to say that viewing an animal as a desiring subject requires

---

[18]It is worth noting that it would be possible to expand the structure of desire further, e.g., by including the time of the desiring and the situation the desiring subject is in as elements, but this would be rather superfluous in the present context.

[19]That is, we do not attribute desires to human beings *qua* complex biological systems that display *behaviour* of various kinds. To view a human being (or some other animal) from "the perspective of natural science" is (on my view) incompatible with attributing desires to it. (Viewed purely as a biological system, an animal may have needs (of a physiological kind), but it cannot have any desires).

that we *first* view it as a conscious agent, however: To attribute full-fledged desires to an animal *is* to view it from a psychological perspective, to view it as a desiring subject is *part of* viewing it as a conscious agent. (But to attribute a *particular* desire to it seems to require that it is first viewed as a conscious agent, i.e. as a desiring subject). The point is: "desire-hood" cannot plausibly be regarded as separate from consciousness or voluntary agency, the three stand and fall together, so to speak[20] [21].

The object of desire: *desire for* and *desire that*

As said, to want is always to want something; desires are intentional states, they are directed toward objects. But what is the object of desire? When someone wants something, what kind of thing is it that is being wanted?

There are two main types of answer that can be given to this question, viz. the *thing-view* and the *situation-view*. On the former view, the objects of desire are things (or things-under-descriptions), and to desire something is to have a *desire for* such a thing. On the situation-view, the objects of desire are situations (or situations-under-descriptions), and to desire something is to *desire that* some situation obtains,

---

[20]Cf. Kenny's (1989) view on the connection between desiring and consciousness. According to him, "consciousness cannot be identified as such independently of wanting. /.../ [T]he notions of consciousness and wanting become applicable together when the behaviour of an agency manifests the appropriate degree of complexity" (p 34). As far as the relation between being a desiring subject and being an agent (between desire and action) is concerned; see below.

[21]Is there anything else that can be generally said about human beings *qua* desiring creatures? Well, it is sometimes claimed (e.g., by Nussbaum (1986), Ch. 9, and Kenny (1989), Ch. 3) that every desiring creature is a needy creature, and we would not have desires if we did not have needs. This might well be true, but there is (as I see it) no reason why we should regard the alleged connection between need and desire as conceptual. First, it is pretty obvious that there are particular desires which are not conceptually connected to particular needs. (This is so, even if one falsely assumes that "desiring X" implies "lacking X"). Second, if the idea is that the general notion of desire somehow presupposes the general notion of need, I can't see how. There are at least some kinds of desiring which seem to be fully compatible with not needing anything at all. God, for example, is sometimes conceived of as a being with plans, goals and purposes (i.e. as an agent), but without needs. This may be a silly conception of God, but it is surely not a conceptual mistake! In short, we should conceive of the connection between need and desire as contingent rather than conceptual.

i.e. *that* something is the case[22].

So, which of the two views is the most ethically relevant view of the object of desire, and which of the two views should the desire theorist accept? Before we look at this issue, let us first take a closer look at the two views.

The thing-view

In ordinary language, desire-statements are sometimes of the form "P (a desiring subject) desires X" or "P has a desire for X" (rather than "P desires that X" or "P desires to Φ"). Examples of this are when we say that someone desires (or longs for, or craves, or has a desire for) a certain woman (or man), a certain job, or a certain car. This *suggests* that the objects of desire are (sometimes) things other than situations (but cf. below).

The thing-view is often combined with the idea that desiring X implies lacking X; to desire something, or want something, is to be *in want* of it (cf. Jeffrey (1983), pp 62-63). It is not surprising that the two views are so easy to combine. The kind of desire which seems to best fit the thing-view is unsatisfied appetite (longing, or craving); "to have a desire for X" often means "to have an unsatisfied appetite for X". And since this prototypical kind of *desire for* is essentially unsatisfied, it is easy to see how well the thing-view fits the idea that "to want is to be in want of".

Nussbaum (1986) seems to attribute the thing-view to Aristotle, when she characterizes *orexis* (Aristotle's general notion of wanting or desiring) as a selective, object-oriented, active "going for", or "reaching after", objects that are seen to have a certain relation to one's needs (pp 274-276). What she seems to suggest is that *orexis* should be conceived of as *orexis for* something, and that the objects of *orexis* are things rather than situations.

The situation-view: Desires as propositional attitudes

The situation-view seems to start from the idea that the typical desire-statement is of the form "P desires that X is the case", or that it can

---

[22]It should be noted that this difference between "desire for" and "desire that" seems to be pretty analogous to the difference between "thinking of" and "thinking that".

easily be translated into a statement of this form[23]. Examples of such statements are "I want you to do this for me" and "I desire that the weather will be fine tomorrow". This suggests that the typical object of a desire is a *situation*, i.e. a state of affairs, an event, or the like. To desire is to *desire that* something is (or will be) the case, that some situation holds or obtains (will hold or obtain)[24].

Because of the intimate connection between situations and propositions, the situation-view can also be expressed in "propositional terms" (this is why desires are, on the situation-view, regarded as propositional attitudes). *Qua* intentional states, desires (this goes for *desires for* as well as for *desires that*) have "semantic content", they require (as Hare (1981) puts it) "a linguistic form for their full description". Now, the semantic content of a *desire that* is always a proposition: When someone desires that something is the case, we always describe this by saying that he desires *that* X, where X is a proposition (and where "X" is a sentence). To desire that X is, so to speak, to have a particular attitude toward the proposition X. This does not imply that the objects of *desires that* are propositions, however. To *desire that X* is, rather, to desire that the proposition X is *true*, i.e. it is also possible to think of the object of a *desire that* as the truth of some proposition.

The situations which are the objects of our *desires that* are probably, in most cases, more complex than we think. The propositional contents of these desires are rarely as simple as "state of affairs S holds", "person P is in mental state M", "event E happens", or "person P performs action A". Suppose I want you to help me move my furniture. This is not a full description of what I want. A fuller specification of the object of my desire would have to include things like "and I want you to help me before Sunday" and "I don't want you to help me if you don't want to". And if I desire that Z. wants me, I also want her to want me at some time or times (e.g., "now" or "forever" or "not all the time"), in

---

[23]In ordinary language desire-statements are often of the form "P desires to Φ", where "Φ" is a verb phrase. Examples of such statements are "P desires to win" or "P wants to go for a swim". However, these desire-to-statements are easily reformulated as "P desires *that* he wins" or "P desires *that* he goes for a swim".

[24]Past situations are normally not objects of desire. An interesting issue, where different opinions are held, is whether past situations *can* be objects of desire. We sometimes say things like "I wish it wouldn't have happened", but are such utterances really expressions of desire? We will return to this question below, in section 4.3.

the right way (e.g., "not too possessively"), and for the right causes (e.g., "freely, and not because she was given an aphrodisiac")[25].

Why the situation-view is superior in ethical contexts

At this point, it should be rather obvious that the desire theorist should accept the situation-view. The reason for this is simple: Because of the intimate connection that holds between the object of a *desire that* and its fulfilment, we know what it is for a *desire that* to be fulfilled: a desire that X is (roughly speaking) satisfied if and only if X actually holds. In the case of *desire for*, on the other hand, there is no such direct connection between object and fulfilment, and this makes it hard to tell what the fulfilment of such a desire consists in. In fact, the only reason why we can often determine whether a certain *desire for* has been fulfilled is that most *desires for* correspond to a number of *desires that*. Suppose a person P wants, has a desire for, a certain woman. How can we tell (or know) whether this "appetite" of his is fulfilled or not? It should be noted that the disappearance of the desire is not a sufficient condition for fulfilment. The fulfilment of P's desire may, of course, cause it to disappear, but it may also disappear in other ("wrong") ways, e.g., because he met someone else, or because time passed, or because he took a desire-eliminating pill. So, how do we distinguish between the fulfilment of a *desire for* and the disappearance of a *desire for*? For this to be possible, there must be some kind of correspondence between desire-for-statements and desire-that-statements, i.e. desire-for-statements must, at least in some approximate way, entail desire-that-statements.

Now, luckily enough, this is often the case, e.g., if P has a desire for a certain woman, this seems to *entail* (in some sense of the term) that he desires that a number of situations obtain (or will obtain), situations in which the woman in question is a "component part". In general, to have a desire for something seems to be connected with things like desiring to possess it, own it, or consume it, wanting it to be present, or desiring to be in contact with it. As Jeffrey (1983) points out, "in such cases [when the objects of desires are not propositions] the highly flexible notion of having something is linked with the notion of desiring it: to

---

[25]This suggests that many of our complex desires can be fulfilled to different degrees (at least in those cases where the propositional content of the desire is naturally conceived of as a conjunction).

desire x is to desire that one have x (in the appropriate sense of 'have')" (p 60).

In short, the only reason why we understand what it is for a certain *desire for* to be fulfilled is that it is (in a certain way) connected to a number of *desires that*. It might even be suspected that *desires for* cannot (properly speaking) be fulfilled at all (at least not "as such"). For example, if one has a clear idea of what would constitute the fulfilment of what seems to be a particular *desire for*, then this desire is probably a *desire that* rather than a *desire for*.

What we have shown this far is that *desires for* can not be directly relevant in a desire theoretical context; the only way in which they can be of relevance is indirectly, by corresponding to a number of *desires that* which are themselves of direct relevance. So, is there any other way in which a person's *desires for* can be directly relevant for his well-being? I think not, but there is at least one other way in which they may be of indirect relevance, viz. for hedonistic reasons[26]. The perhaps most typical example of *desire for* is, as we have seen, unsatisfied appetite. If someone has an unsatisfied appetite for something, this is often something he feels, and the feeling in question is normally unpleasant. Furthermore, an unsatisfied appetite is often (especially if it is a bodily appetite) felt more or less continuously until it is satisfied. It is, I think, mainly the connection between unsatisfied appetite and unpleasantness which makes it ethically relevant that someone has an unsatisfied appetite for something, but we should also notice that it is often pleasant to have a such desire satisfied.

To conclude, the ethical relevance of *desires for* either consists in the fact that they may be unpleasant to have, or in the fact that they may correspond to a number of *desires that* (which may be, in themselves, ethically relevant). That is, the notion of *desire for a thing* is not directly relevant in this context, and a plausible conception of well-being do not have to take our *desires for things* into account.

---

[26]*Desires that* do (of course) not differ from *desires for* in this respect, i.e. they may also be indirectly relevant in this (hedonistic) way.

The desiring (by the subject, of the object): A rudimentary conception of desire

What is it for someone to desire that something is the case? How, for example, does the desire that X obtains differ from other propositional attitudes, like the belief that X obtains? What conception of desire (i.e. *desire that*) is it likely that the desire theorist has in mind, and above all, what conception of desire (and aversion) makes the desire theory most plausible?

On my view, there are a number of conditions that a conception of desire (and aversion) must satisfy if it is to qualify as maximally relevant in this context, viz. the following ones:

First of all, the most relevant conception of desire must accept the following two claims:

(i) Desires and aversions are propositional attitudes (cf. above).

(ii) Desires are pro-attitudes, while aversions are con-attitudes, no matter how the contents (objects) of these attitudes are specified[27]. This idea can be given a more exact formulation in terms of preference and indifference, viz. in the following way. Let us first define "neutrality" as follows: A person P is neutral to the occurrence of a certain situation X if and only it does not matter to him whether X holds or not, i.e. if and only if he is indifferent between X and not-X (which implies that he does not prefer X to not-X, or vice versa). Armed with this definition of neutrality, we can now say that P *desires* that a situation X obtains if and only if he prefers X to some situation to which he is neutral, and that P has an *aversion* to a situation Y if and only if he prefers some neutral situation to Y[28].

(iii) Moreover, we can also assume that the desire theory cannot be plausible unless the desire theorist uses the term "desire" in a very

---

[27]It is worth noting that this is inconsistent with the (implausible) idea that the difference between desires and aversions is a difference in content (or object), viz. that the content of a desire is positively specified, while the content of an aversion is negatively specified, i.e. that an aversion is (by definition) a desire that a certain (positively specified) situation does *not* obtain.

[28]This has certain interesting implications, viz. the following ones: Suppose that P prefers X to not-X. In this case, we can not conclude that P desires that X (in the above sense), i.e. we can not conclude that it is good for P that X obtains. Neither can we conclude that P has an aversion to not-X, i.e. it is not necessarily the case that not-X is bad for P. The only thing we can conclude from the fact that P prefers X to not-X is that *either* X is good for P *or* not-X is bad for P.

broad sense, i.e. unless he uses the term in such a way that a very large and heterogeneous class of mental states can be subsumed under it. To get an idea of how broad the relevant notion of desire may be, consider the following quotations. The first one is from Kekes (1988):

> Some of them [our desires, or wants] are insistent enough, so there is no special difficulty in being aware of them. But not all wants speak with a loud voice; some are inarticulate; some concern the distant future; some have disturbing emotional undercurrents; some are confused longings for impossible goals, like wishing the past to be different, or undoing present unpleasantness in fantasy, or hoping for wildly unrealistic future developments (p 159).

The next quotation is from Kenny (1989). In his terminology, there are two fundamental kinds of want, viz. "desires" and "volitions". This is how he characterizes the difference between the two:

> /.../ desire, unlike volition, seeks immediate satisfaction; that is, it is a want for something now, a want that is felt more or less continuously until it is satisfied. Volition, by contrast, may be for something distant in time, and it may be operative without being an item in the flow of consciousness (p 36).

He also adds (on p 37) that "volition involves the exercise of concepts which need language for their expression, whereas desire need involve only the exercise of simpler and more rudimentary concepts, which can be manifested in non-linguistic behaviour".

So, it seems that on the relevant use of "desire", the class of desire includes things as different as volitions and intentions, appetites and longings, projects and purposes, requirements and demands, wishes and regrets, i.e. the class of desire is, in many respects, a very heterogeneous class. So, in virtue of what common feature (assuming that there is such a feature) do all these different psychological states belong to the same category, i.e. the class of desire? This is a question that a more complete conception of desire has to answer. (However, it is likely that the grouping of all these states into a single genus "involves a degree of philosophical regimentation" (cf. Kenny (1989), p 41), i.e. it can be suspected that the relevant notion of desire is a rather technical notion).

(iv) It is also desirable that a conception of desire does not deviate

too much from the ordinary use (or uses) of the term "desire" (cf. p 160 above), especially not from its "explanatory" or "theoretical" use (i.e. the use from which the technical, philosophical notion of desire has "arisen"). We often use the term "desire" in this theoretical sense, e.g., when we say that all actions are (in part) manifestations of desires, or that all actions must be explained in terms of desires and beliefs, or that "a desire may be had in the absence of its being felt" (cf. Smith (1994), p 109), or that we do not always know what we want (that we are sometimes fallible about the desires we have). It may not be necessary that a relevant conception of desire is consistent with *all* the ordinary uses of the term "desire", however (since it might simply not be possible to construct such a conception[29]).

Now, it is easy to see that this rudimentary conception is far from complete. For example, we do not really know what an attitude is, or what it is for someone to prefer something to something else[30]. So the question arises: How should we complete the rudimentary conception above, i.e. what must a complete conception of desire be like if it is to satisfy conditions (i)-(iv)? For example, should we adopt some kind of phenomenological conception of desire, or should we (instead) go for a functional conception? This question will not be discussed here, however, but in appendix F.

4.2. What is it for a desire to be stronger than another desire?

So, when the desire theorist claims that the value it has for a person to have a relevant desire fulfilled is a function of one thing only, viz. how strong the desire is, what conception of strength should he have in mind? What conception of strength makes the theory most plausible?

As far as I can see, there are (in this particular context) three ways in which the term "strength" can be understood, viz. (i) as felt intensity, (ii) as motivational force, or (iii) as rank in a preference ordering (where preference is *not* understood in terms of felt intensity or motivational

[29]For example, it might not be possible to construct a conception of desire that is consistent both with (i) the idea that we are sometimes directly aware of our desires, and (ii) the idea that all actions are (in part) manifestations of desires. Cf. appendix F.

[30]The notion of preference will be treated more in detail in section 4.2.

force)[31]. Let us take a closer look at these three senses of the term, and see which one that makes the desire theory most plausible.

(i) On the first sense of "strength", a desire is stronger than another desire if and only if it is more intensely felt[32]. This can not be the relevant sense of "strength", however. First, it presupposes (falsely) that some kind of phenomenological conception of desire is true (cf. appendix F), and second, there is no reason to assume that it is better for a person to have a desire fulfilled just because it is more intensely felt. But why is this? On Griffin's (1986) view, "felt intensity is too often a mark of such relatively superficial matters as convention or training to be a reliable sign of anything as deep as well-being" (p 15), but doesn't this hold for the other two senses of "strength" as well? On my view, the reason is (rather) that the felt intensity view makes the strengths of our desires too dependent of our probability judgements (cf. below).

(ii) If "strength" is understood in the second way, the "strongest" desire is "the motivational winner", the desire with the strongest motivational force[33]. This is probably the sense of "strength" that is most in harmony with the functional conception of desire (cf. appendix F), especially with the idea that desire is an explanatory notion, but this

---

[31]There is also a fourth sense, viz. "rank in a cool preference ordering, an ordering that reflects appreciation of the nature of the objects of desire" (Griffin (1986), p 15), or "place in an informed preference order" (cf. ibid., p 99). On Griffin's view, this is "the relevant sense of 'strength'", the sense that is incorporated into Griffin's own "informed-desire account of well-being". This is not a conception of strength that the *unrestricted* actual desire theorist can accept, however.

[32]There are (it seems) two possible interpretations of the claim that a desire D1 is more intensely felt than another desire D2, viz. (a) D1, or some sensation (e.g., tension) which accompanies D1, is more strongly felt, and (b) the object of D1 exerts (phenomenologically speaking) a stronger attraction, or pull, on the subject.

[33]This notion of strength is probably identical with what Gauthier (1986) calls revealed preference, and what Sumner (1996) calls the behavioural notion of utility, or preference in terms of actual (market) choice, i.e. the view that a person prefers X to Y just in case he actually chooses X when he could have chosen Y instead (whatever *that* means). This behavioural notion of preference is contrasted against the attitudinal notion (which is, on Sumner's view, the relevant notion in this context), according to which a person prefers X to Y if and only if he likes X better than Y, or finds X more agreeable. But what is *this* supposed to mean, i.e. how should the term "attitude" be understood in this context (on the assumption that it should not be understood in behavioural terms)? Well, as far as I can see, there are only two possibilities here: Either the attitudinal view is just another version of the "felt intensity" view described above, or we have to think of attitudes in functional (but not behavioural) terms.

does not make it relevant in the context of well-being. As Griffin writes, "the relevant sense of 'strength' is not simply the desire that wins out in motivation", because if "strength were interpreted as motivational force, then 'utility' would lose it links with well-being" (ibid., p 15). But why? Well, first, the motivational force view makes strength far too dependent of probability judgements (even more so than the felt intensity view), and second, the view suggests that to the extent that our actions are successful, and to the extent that the desires we act on are intrinsic, we are as well off as we can be, but this is surely an implausible view[34].

(iii) That a person's desire that X is stronger than his desire that Y in the third sense means that he prefers X to Y. This idea can also be formulated in terms of *utility*, viz. in the following way: A person P's desire that X is stronger than his desire that Y if and only if X has higher *utility* for P than Y has (if X is more desirable for P than Y is)[35].

But how should this idea be interpreted, i.e. what is it for someone to prefer something to something else? Or more specifically: If we want to keep the preference view distinct from both the felt intensity view and the motivational force view, what conception of preference is most relevant in this context? Well, personally I tend to accept the idea that preference should be understood in terms of hypothetical choice, i.e. that a person prefers X to Y if and only if he would, under conditions of certainty, choose X. However, it has also been suggested (e.g., by Gauthier (1986), p 27) that "attitudinal preference" should be under-

---

[34]Another problem with the behavioural view (which was pointed out to me by Jan Österberg) is that it can not account for the difference between picking and choosing, where "[w]e speak of *choosing* among alternatives when the act of taking (doing) one of them is determined by the differences in one's preferences over them. When preferences are completely symmetrical, where one is strictly indifferent with regard to the alternatives, we shall refer to the act of taking (doing) on of them as an act of *picking*" (Ullmann-Margalit and Morgenbesser (1977), p 757).

[35]Where utility is (roughly speaking) essentially a reflection of, or measure of, a person's preferences; in this case, his *intrinsic* preferences. For example, to say that X has more utility for P than Y has is equivalent to saying that P prefers X to Y, and to say that X and Y has the same utility for P is equivalent to saying that P is indifferent between X and Y. (And here, it doesn't really matter which of the two possible interpretations of "utility" one has in mind; utility as subjective desirability (or relative importance to the desiring subject), or utility as a *measure* of subjective desirability (i.e. a *number* that reflects the relative importance of an outcome)).

stood as something which is "essentially" expressed in speech, i.e. verbally. But can (and should) this view really be regarded as a claim about what preference *is*, e.g., should we attribute to Gauthier the idea that a person prefers X to Y if and only if he would (when asked) express (in speech) a preference for X over Y[36]? I think not. On my view, the idea is better understood as an idea of how we should gain knowledge of someone's preferences[37].

Regardless of exactly how the notion of preference is to be understood, it is (I think) pretty clear that "rank in a preference ordering" is the sense of "strength" (of the senses that are, so to speak, "open" to the unrestricted desire theorist) that is most closely linked with well-being, and one important reason for this is that it is the use of "strength" on which the strengths of our desires are most likely to be independent of our probability judgements (but cf. e.g., pp 208-211 below). The felt intensity of a person's desire that X is most certainly dependent on how probable he thinks it is that X occurs (or can be realized), and the same thing holds (to an even larger extent) for the motivational force of the desire. But it is surely implausible to make the value that a certain situation has for a person dependent on his probability judgements. And since "rank in a preference ordering" is (it seems) also most consistent with the rudimentary conception of desire in section 4.1 above, this is the sense of "strength" that the unrestricted desire theorist should accept[38].

Or alternatively put, the interpretation of (UD3) that makes this claim most plausible is this: It is nonderivatively better for a person to have his intrinsic desire that X (or aversion to X) fulfilled than to have his intrinsic desire that Y (or aversion to Y) fulfilled if and only if he intrinsically prefers X to Y, i.e. if and only if X is ranked higher than Y on his intrinsic preference ordering.

---

[36]But what question should he be asked? Here it might be argued that we should ask the person what he would choose (but not pick) under conditions of certainty, i.e. the two ideas might not be so different after all.

[37]Another possible view on the connection between preference and choice is this: If a person prefers X to Y, then he has a non-arbitrary tendency to choose X instead of Y (cf. Bergström (1991), p 10).

[38]It is also worth mentioning that this third conception of strength is also the most methodologically appealing one. Or more specifically, it is the conception that makes strength intrapersonally and intratemporally measurable to the highest degree. However, if we have the interpersonal or (intrapersonal) intertemporal cases in mind, it is not as appealing.

4.3. Desire and time: Some temporal issues

We have already seen (on pp 161-162) that on the traditional use of "desire-fulfilment", the time of the desire need not coincide with the time of its fulfilment (i.e. the time of its object). Or more specifically, if this notion of fulfilment is adopted, both "now-for-then desires" (desires about the future, or "prospective desires") and desires about the past (or "retrospective" desires)[39] can (assuming that there is such a thing as full-fledged intrinsic desires about the past) be fulfilled. So the question arises: Should the unrestricted desire theorist accept this broad notion of fulfilment, i.e. should he accept (a) the idea that it can have nonderivative value for a person to have his "prospective desires" fulfilled, and (b) that it can (also) be good for a person to have his "retrospective desires" fulfilled?[40]

(a) So, is it nonderivatively good for a person to have an intrinsic now-for-then desire fulfilled? For example, is it possible for a person to improve his life by fulfilling the prospective desires he once had but no longer has? And can it be good for a person that his present prospective desire will be fulfilled in the future, i.e. can a person's well-being be affected "retroactively"? I think not, and for the following reason: Suppose that a person P has an intrinsic desire at t_1 that a situation X holds at some future time t_2, and that it is claimed that it is nonderivatively good for P to have this desire fulfilled. We can then ask:

---

[39]The term "now-for-then desire" is from Hare (1981), and the terms "prospective" and "retrospective" from Bykvist (forthcoming).

[40]It is worth noting that all this is inconsistent with Sumner's (1996) idea that desires are essentially prospective (cf. pp 128-129). This is not the only way in which Sumner's conception of desire differs from the conception adopted here, however. For example, desire has (on Sumner's view) no negative counterpart; according to him, "[a]version is the opposite not of wanting but of liking"; cf. note 24 on p 125. Moreover, desire seems (on his view) to imply lack; cf. p 129, where he writes that "I can want only what I have not yet got" (but shouldn't he write "what I *believe* I have not yet got"?). In short, Sumner's notion of desire is quite narrow, something which is intimately connected with the view that desiring X is only one of many possible ways in which we can be in favour of X (where liking X, enjoying X, approving of X, and endorsing X are examples of other possible ways). On the conception adopted here, desire is a much broader phenomenon (e.g., most likings are regarded as desirings). However, this (broader) notion of desire is not necessarily as broad as the notion of pro-attitude; there *may* be positive attitudes which can not be regarded as desires, e.g., enjoyments (where enjoying X implies being aware of X).

For whom is it supposed to be good that X obtains at t_2? Good for P, of course, but good for P-*when*? It seems that it cannot be good for P-at-t_2 (when he no longer has the desire)[41]. Neither can the fact (if it is a fact) that X *will* occur at t_2 be ("retroactively") good for P-at-t_1. This suggests that there is no point in time t at which it is good for P-at-t to have the desire fulfilled. So, can it be good for P-over-time that X obtains at t_2? I think not. On my view (cf. appendix C), all goodness-for-P is goodness-for-P-at-some-time, and there is (for this reason) no such thing as goodness-for-P-over-time. And if it is also the case that only things which occur at t can be nonderivatively good for P-at-t, this means that the unrestricted desire theorist can (and should) regard all prospective desires as irrelevant[42].

If this is so, it can never be good for a dead person P to have his last (prospective) wish fulfilled (assuming that death is real, that is). If death means annihilation, there is no such thing as P-after-death (for whom things can be good or bad), and neither can it be "retroactively good" for P-before-he-died to have the wish fulfilled. And if there is no point in time t at which it is good for P-at-t to have the wish fulfilled, then it can not be good for P to have it fulfilled (this follows from the idea that

---

[41]Especially not if P-at-t_2 is not aware of the fact that the once desired situation obtains, or if "the desired state of affairs turns out upon later experience to be disappointing or unrewarding" (Sumner (1996), p 132). In this context, it is also worth noting that "just as we can be disappointed when we get what we expect or have aimed for, so we can be pleasantly surprised when we get what we do not expect and have not aimed for" (ibid., p 132). That is, a situation may well be good for a person even if he had no prospective desire for this situation, and this means that *if* desires were essentially prospective (as Sumner believes), then (D1) would be refuted.

[42]However, it might be objected: "Of course it can be good for me to have my past desires fulfilled. This idea can only be rejected by someone who has an extremely superficial conception of what it means to have an identity. The connection between who and what I am now, and who and what I once used to be, is not merely a causal connection; the connection is much stronger than that. We are 'historical beings', e.g., what we are is not just causally dependent on our respective histories, it is also (in part) constituted by how we conceive of our respective histories. My 'life story' (as I conceive of it) is (somehow) part of my identity, and this means that 'my past selves' are (in a sense) 'parts' of my 'present self'. This implies that some of my past desires may (in some sense) be present desires, especially if they once played an important and central role in my life. For this reason, it may be good for me to have them fulfilled now; this may, for example, improve my story!". The argument is interesting, but I still tend to believe that the fulfilment of past prospective desires can not contribute directly to a person's well-being.

all goodness-for-P is goodness-for-P-at-some-time). This is not to say that it can never be *good-period* that a person's last wish is respected, however. In fact, I think we often have a reason to try to satisfy the prospective wishes of the dead, regardless of whether this is (strictly speaking) good for them or not.

(b) The next question is whether it can ever be nonderivatively good for a person to have an intrinsic retrospective desire (a desire about the past) fulfilled[43]. Suppose P has an intrinsic desire at t_2 that a situation X obtained at some past time t_1 (e.g., that he desires now that he had a job last year). Is it good for P to have this desire fulfilled (that he, in fact, had a job at the time)? I think not. It is not good for P-at-t_1 (who does not yet have the desire), and neither is it (or so it seems) good for P-at-t_2, nor for P-over-time. In short, if all goodness-for-P is goodness-for-P-at-some-time (and if X can not be nonderivatively good for P-at-t unless it occurs at t), it can not have nonderivative value for us to have our retrospective desires fulfilled. Moreover, the idea that it has nonderivative value for us to have our intrinsic retrospective desires fulfilled implies that we can improve the quality of our present lives *directly* (and to a considerable extent) simply by wanting (intrinsically) that our

[43]Assuming that there are such desires, that is. But is this really so? That is, (i) are there such things as full-fledged retrospective desires, and (ii) if there are, are these desires ever intrinsic? (i) Well, it seems that there are full-fledged retrospective desires; my regretful wish that I behaved differently yesterday is (it seems) a good example of a such a desire. This common sense belief is not necessarily consistent with a functional conception of desire, however (cf. appendix F). It is true that my retrospective wish may well be connected to action-tendencies and/or dispositions to act in certain ways (e.g., the tendency to compensate the person I wronged, or "to set things right"), but (so the argument goes) the wish is not *properly connected* to action-tendencies. The tendency to "set things right" is properly connected to the *prospective* desire to "set things right", a desire that may or may not be caused by my retrospective wish. It might be objected, however: "The reason why our retrospective desires are not 'properly connected to action-tendencies' is not that they are retrospective, but that they are impossible to satisfy (and that we believe this), and that they are (for this reason) impossible to act on. If we could affect the past, and if we would believe that it were possible (e.g., by making time travels, or by asking God to change the past), then we would sometimes take steps to undo parts of the past, and in these cases our retrospective desires would be properly connected to action-tendencies. This means that we can accept the functional conception of desire and still regard these desires as full-fledged desires". (ii) So, are retrospective desires ever intrinsic? Well, it seems possible that they are, but it is (I think) rather "abnormal" to have such desires.

respective histories are what they are, and this is absurd[44]. That is, the unrestricted desire theorist can (and should) regard all retrospective desires as irrelevant.

To sum up, the unrestricted desire theorist should regard both prospective and retrospective desires as irrelevant, That is, he should accept the idea that if the time of a desire does not coincide with the time of its object, i.e. with the time of its fulfilment (in the traditional sense), then it can never have nonderivative value for the desiring subject to have the desire fulfilled. Or alternatively put, the desire theorist should reject the traditional notion of fulfilment and replace it with the idea that P's desire that X is fulfilled (in the relevant sense) if and only if P desires that X, X holds, *and* the desire and its object are simultaneous (the time of the desire coincides with the time of the occurrence of X)[45] (cf. (i) on p 162).

Now, if we accept this view, there are (it seems) a number of problems that we will have to face. The first problem is this: The synchronist view seems to suggest that Bert's present desire to be together with Alice from now on (e.g., until death do them apart) and his present desire to be together with Alice now (where this now has at least some temporal extension) are equally relevant, i.e. that it is equally good for Bert-now to have the two desires fulfilled (in the traditional sense)[46]. But is this really a plausible idea? Well, this seems to depend on how the propositional content of the latter ("now-for-nowish") desire is specified. If the object of this desire is to be together with Alice now (period), I don't think the claim is plausible. However, if the object of the desire is (instead) to be together with Alice now "*qua* part of a much larger temporal whole"[47], the claim is far from implausible: it is,

---

[44]This is (of course) not to say that it can not be good for a person to be satisfied with his past *for other reasons*, e.g., for hedonistic reasons.

[45]This is the view that Tännsjö (forthcoming) calls "The Theory of Simultaneous Satisfaction of Preferences", and it is similar to the view that Bykvist (forthcoming) calls "synchronism". It is also implicit in Hare's (1981) conception of happiness (according to which happiness is a matter of fulfilment of "now-for-now desires"; cf. p 103).

[46]This is based on the assumption that the "now-for-now content" (or "now-for-now component") of the former desire is identical with the content of the latter (now-for-now) desire.

[47]So, when Bert desires to be together with Alice now *qua* part of a much larger temporal whole, what exactly is it he desires? Well, roughly, he desires (i) that he be together with Alice now, (ii) that their relationship is open, that their future prospects are good, and (iii) that this permeates their being together. (This is, I

after all, likely that the situation "being with Alice now *qua* beginning" ranks much higher on Bert's preference ordering than the situation "being with Alice now (period)". But this gives rise to yet another problem: Are desires for things *qua* parts of larger temporal wholes really now-for-now (or "now-for-nowish") desires, i.e. can the synchronistic view really regard such desires as relevant? Well, if the interpretation that was given in note 47 is adopted, I think the answer is "yes". In short, it seems that the synchronist view need not "ignore" the fact that we are future-oriented, i.e. that some of our strongest desires are (in part) prospective[48]. (We may also ask whether desires for things *qua* parts of larger temporal wholes are really intrinsic. I think they are, at least if we take "intrinsic" to mean "final" rather than "intrinsic proper"; cf. pp 183 and 186 below).

This is another problem which is actualized by the synchronistic view: It seems that there are desires (e.g., certain bodily appetites) which disappear when they are satisfied. Moreover, some of these desires (e.g., longing) are "essentially" desires for something which is absent: they are essentially connected to lack. Such a desire can exist only as long as it is not satisfied, and it therefore disappears as soon as it is fulfilled. Now, it seems that it is good for a person to have such a desire fulfilled, even though the time of the desire does not coincide with the time of its fulfilment[49]. Does this mean that we have to reject the synchronistic view? I think not. Or more specifically, I think that if we understand why it is good for us to have such essentially unsatisfied desires fulfilled, we will see that there is no real threat to this view.

So, suppose that P has an (intrinsic) unsatisfied desire (e.g., that he is thirsty), and that this desire is fulfilled at t, and therefore disappears. Suppose also that it is (nonderivatively) good for P to have the desire fulfilled. But why is this? If we accept the sychronistic view, it seems

---

suggest, the "now-for-now component" of Bert's present desire to be together with Alice from now on).

[48]It is worth pointing out that Bykvist (forthcoming) has, for similar reasons, given up the synchronistic view altogether. His own alternative is what he calls "the existence requirement".

[49]This is (of course) an idealization. In real life, there is almost always some overlap in time, e.g., if I am thirsty and strongly want to drink some water, it is unlikely that this desire disappears as soon as I start drinking; it will at least take a few seconds before it disappears. But then it really disappears, i.e. it does not just seem to disappear, it is not just the unpleasant feeling with which it is associated that disappears.

that we cannot explain this intuition on desire theoretical grounds: If the time of the desire does not coincide with the time of its fulfilment, and if all goodness-for-P is goodness-for-P-at-some-*t*, then it can not be good for P to have the desire fulfilled. It is not good for P-before-*t* (who has the desire), and neither it is good for P-after-*t* (who no longer has the desire). So, why is it good for P to have the desire fulfilled? As I see it, there are two possible explanations, both of which are fully compatible with "the theory of simultaneous satisfaction of preferences":

(i) It is good for hedonistic reasons: The reason why this kind of desire-fulfilment is good is not that the desire is fulfilled, but that it is eliminated. And the reason why it is good for P to have it eliminated is mainly that it is unpleasant to have unsatisfied appetites, but it may also be positively pleasant to get rid of them.

(ii) What we think of as an unsatisfied appetite is a complex thing, e.g., to be thirsty is not merely to have a desire to drink, it also contains an aversion to not-drinking, and maybe also an aversion to the feeling of thirst (aversions that may well be intrinsic). What is "good" for P when he has the appetite satisfied is not (strictly speaking) that he has a desire fulfilled, but that aversions that were fulfilled before *t* are no longer fulfilled, i.e. his well-being has been improved by removing something bad rather than adding something good.

In short, the problem concerning "essentially unsatisfied appetites" are not really a problem, and we can retain the "synchronistic view".

The remaining temporal issue has to do with duration rather than with the relation between the time of a desire and the time of its fulfilment. The problem I have in mind is this: Suppose a person P has a desire fulfilled at *t*. Suppose also that the desire came into existence before *t*, i.e. that it has been unfulfilled for some time. Does it (on the desire theory) matter how long P had the desire before it was fulfilled? Or more specifically, if P has two intrinsic desires that are equally strong, and if one of them is "old" (he has had it for a long time) and the other is much "younger" (it came into existence more recently), isn't it (nonderivatively) better for P to have the older desire fulfilled than to have the younger desire fulfilled? I think not. How old a desire is is not in itself relevant; it is relevant only if (and because) P himself prefers to have older desires fulfilled than to have younger desires fulfilled, i.e. if the "age" of a desire will (for P) have an impact on its strength. That is,

the unrestricted desire theorist need not reject the idea that "strength is what matters" (at least not for *this* reason); he can safely accept the idea that it is better for the desiring subject to have the stronger of two intrinsic desires fulfilled, no matter how long this subject has had the desires[50].

4.4. The intrinsicality condition

We have already seen that the desire theorist (as characterized by me) accepts the "intrinsicality condition", i.e. the view that intrinsicality is a necessary condition for relevance, that it can never be nonderivatively good for a person to have a desire fulfilled unless this desire is intrinsic. In order to determine whether this is a plausible condition, we first have to know how it should be understood, i.e. what it is that makes a desire intrinsic.

So, what is it to desire something intrinsically, in an intrinsic way? Let us look at some typical answers that have been given to this question. According to Rabinowicz and Österberg (1996), to prefer something intrinsically is to "prefer it for its own sake" (p 1), and according to Parfit (1984), intrinsic (or "un-derived") desires are desires that are *not* "for what are mere means to the fulfilment of other desires" (p 117). And in Brandt's (1979) terminology, to have an intrinsic desire that a certain situation obtains is (roughly) to simply want it to obtain, "for no further reason", and not because of its probable further effects[51].

---

[50]Another problem concerning duration and value-for is this: If it is nonderivatively good for a person to have his intrinsic desires fulfilled, and nonderivatively bad for him to have his intrinsic aversions fulfilled, doesn't this imply that it is (nonderivatively) better for him the longer his intrinsic desires are fulfilled, and that it is worse for him the longer his intrinsic aversions are fulfilled? Now, this problem has already been discussed in connection with hedonism (cf. appendix E), and my answer is again that such claims do not really make sense. (There is no such thing as goodness-for-P-over-time, and there is no point in time *t* such that it is better for P-at-*t* that a desire is fulfilled for a longer time). What we can meaningfully say is this: (i) "At every point in time *t*, it is nonderivatively good for P-at-*t* to have his intrinsic *t*-for-*t* desires fulfilled and bad for P-at-*t* to have his intrinsic *t*-for-*t* aversions fulfilled", and (ii) "It is better-period if a certain desire is fulfilled for a longer time than if it is fulfilled for a shorter time".

[51]Or to be more specific, this is the definition of "occurrent intrinsic valence" that Brandt (1979) offers (a definition which is based on his particular version of the functional conception of desire): "/.../ an outcome *O* is occurrently *intrinsically* valenced for a person at a time *t* if and only if his central nervous system at *t* is

None of these answers are precise enough, however. As I see it, there at least three ways in which the phrase "X is desired intrinsically" can be understood in this context, viz. the following ones:

(i) A situation is desired intrinsically (in "the proper sense"; cf. appendix A) if and only it is desired for its intrinsic properties, or in isolation, rather than for its relational properties, i.e. *extrinsically*[52].

(ii) A situation is desired intrinsically if and only it is desired *"finally"*, i.e. as an end[53], rather than *instrumentally*, i.e. as a means, or "for the sake of something else". (That is, to have an instrumental desire that a situation obtains is to want it to obtain for a certain kind of reason, i.e. because it is probable that it will have certain effects, or because it will make certain other situations possible).

(iii) A desire is intrinsic if and only it is *un-derived* rather than *derived*, where a desire is underived if and only if it is not derived. So, what is it for a desire to be derived? Derived from what, and how? The idea is that desires can be derived from other (more fundamental) desires, on the one hand, and beliefs, on the other (just like beliefs can be derived from other beliefs). Here are three examples of how a desire can be derived (or "inferred") from a more fundamental desire and a belief:

P desires to engage in some physical activity
P believes that hiking is a physical activity
Therefore[54], P desires to go hiking[55]

---

such that if he judged (thought with belief) that a certain act by him at *t* would tend to bring about *O*, then, even in the absence of any further judgements about *O* (such as its probable further effects) not contained in the concept of *O*, there would be an increase in his tendency to perform the act. We must also require that the *O* said to be intrinsically valenced be stripped down so as not to include elements irrelevant to the increase of the person's tendency to perform the act" (p 32). Apart from being too imprecise with regard to what it is for a desire to be intrinsic, this definition also suffers from another defect, viz. it implies that a person can only desire something intrinsically if he believes that he can bring it about (which is surely a strange implication).

[52]It is worth noting that this terminology is really easier to grasp in the case of desire-for (where the objects are things) than in the case of desire-that (where the objects are situations). Cf. appendix A, note 16.

[53]This is the second possible interpretation of "for its own sake".

[54]This is (of course) not the whole story, much more than this is required, e.g., that P thinks that hiking is (at the moment) the funniest, or most accessible (to him), physical activity. This holds for the other two "inferences" as well: the conclusions do not follow from the premises unless a number of assumptions are added.

[55]It should be noted that in cases like this, when a desire is derived in this way, a

P desires that his life be good
P believes that having friends is an essential part of a good life
Therefore, P desires that he has friends

P desires to be happy
P believes that if he gets a job, then he will be happy
Therefore, P desires to get a job

These examples are illustrations of what I think are the three major kinds of derivations of desires. The reason why the three inferences are so different is that the beliefs involved (i.e. the propositional contents of these beliefs) are of different kinds: In the first case, the belief is a "subsumption" (or classification); hiking is classified as a physical activity. In the second case, it is believed that something is a part of something else (a whole). In the third case, the belief is (so to speak) "instrumental": it is a belief about what causal connections hold in the world.

So, this is (roughly) how desires *can* be derived. But what is it for an actual desire to *be* derived? Well, in my terminology, a person P's desire DD is actually derived if and only if P has a more fundamental desire D and a belief B such that (i) DD is derivable from D and B, and (ii) P has actually performed the inference (correctly)[56].

Let us now look at how these three distinctions are related to each other. First, what is the relation between the intrinsic/extrinsic distinction and the final/instrumental distinction? Well, to begin with, it is clear that all instrumental desires are extrinsic: To desire X because of its probable further effects is to desire it for its (probable) relational properties, i.e. extrinsically. But is it also the case that all extrinsic desires are instrumental? Not necessarily. Consider the desire to have friends, where this circumstance is not regarded in isolation, but *qua* part of a certain kind of life. This desire is not intrinsic, but it may well be "final"[57]. That is, the first two distinctions are not necessarily extensionally equivalent with each other.

---

fulfilment of the derived desire is also (on the assumption that the belief is true) a fulfilment of the desire from which it is derived.

[56]Or something of the sort: it is possible that (ii) is too strong a condition.

[57]Another possible example of such a desire is the desire to make more money than one's friends. This desire may well be "final", but can it be intrinsic? Well, I tend to think it can, but I am not sure.

So, how is the distinction between underived and derived desire related to the other two distinctions? Well, it seems that all our underived desires are "final" desires (intrinsic or extrinsic)[58], but not vice versa, i.e. there are (obviously) final desires (e.g., intrinsic desires) which are derived. For example, a desire that is derived from an intrinsic desire and a non-instrumental belief of the identification type may well be intrinsic, and the same thing holds for those of our desires that are derived from an intrinsic desire and a non-instrumental belief of the part-whole type[59]. (Examples of this are already given above; we just have to assume that the more fundamental desires in the examples are intrinsic). That is, not all derived desires are extrinsic (e.g., instrumental); there are desires that are both intrinsic and derived.

Let us now ask ourselves which of these three senses of "intrinsic desire" that a desire theorist should accept, i.e. which of the three senses that makes the intrinsicality condition most plausible. For example, should the desire theorist just ignore a person's instrumental desires, or should he also ignore those of his derived desires and/or extrinsic desires which are not (strictly speaking) instrumental?

Let us first establish that the desire theorist should *not* conceive of intrinsic desires as underived desires, i.e. that in order to be plausible, the intrinsicality condition should *not* be conceived of as a condition of "underivedness". To establish this, we have to consider a desire that is (at the same time) intrinsic (in the proper sense), final, and derived, and then ask whether it can (on a plausible desire theory) be nonderivatively good for a person to have such a desire fulfilled. And of course it can (would anyone who had the difference between the two distinctions pointed out to him reject this view?)[60]. It is not entirely unproble-

---

[58]That is, there are no underived instrumental desires; all instrumental desires are derived (e.g., from intrinsic desires and instrumental beliefs).

[59]Desires derived in these ways *may* be intrinsic, but they *need not* be intrinsic: It is not the intrinsicality (or "finality") of a desire that is derived, but its propositional content. For example, that someone wants intrinsically that a certain whole is the case does not imply that he wants all the parts intrinsically (as far as I can see, it doesn't even imply that he *wants* all the parts!).

[60]This difference between the intrinsicality condition interpreted as a condition of "underivedness" and the other interpretations of the intrinsicality condition might be of little practical importance, however, viz. for the following reason: It is often the case that the fulfilment of a derived intrinsic desire is, at the same time, a fulfilment of the intrinsic desire from which it is derived (e.g., if the desire to go hiking is fulfilled, so is the desire to engage in physical activity; cf. note 55 above). But in cases like this, it hardly has nonderivative value for a person to have his

matic to attribute nonderivative value to the fulfilment of derived desires, however. Suppose a person P has two intrinsic desires, D1 and D2, and that D2 is derived from D1. Suppose also that if D2 is fulfilled, so is D1. Is it, in this case, really plausible to attribute nonderivative value-for-P both to the fulfilment of D1 and to the fulfilment of D2? Well, I can't see why not. It is important that we do not conceive of the two fulfilments as separate, however, and that we do not think that both fulfilments contribute to P's overall well-being (cf. pp 158-159 above)[61].

So, which "condition of intrinsicality" should the desire theorist accept; should he accept (i) the idea that it can never be nonderivatively good for a person to have a desire that X fulfilled unless X is desired for its intrinsic properties (in isolation), or (ii) the idea that it can never be nonderivatively good for a person to have a desire that X fulfilled unless X is desired as an end?

Well, on my view, it is rather obvious that the desire theorist should accept (ii), especially if the synchronistic view has already been adopted. Let us return to the example that was given on p 179. If the desire theorist wants to retain the difference between Bert's intrinsic desire to be together with Alice now *qua* part of a life-long relationship (in the sense this was given in note 47) and his intrinsic desire to be to with Alice now (period), and if he wants to give more weight to the former than to the latter (as he should), then he must accept (ii) rather than (i). What matters is what we desire as ends, and not what we desire in isolation.

The unrestricted desire theory has now been given a sufficiently precise formulation, and we can now move on to "the questions of plausibility".

---

derived intrinsic desires fulfilled *because* these fulfilments are, at the same time, fulfilments of some *underived* intrinsic desire.

[61]It is also worth noting that if we regard certain derived desires as relevant, this might (as Jan Österberg has pointed out) mean that there are relevant desires the strengths of which are not as independent of probability judgements as I would want them to be (cf. pp 172-175 above).

Chapter Five

Is the desire theory (in any of its versions) a plausible theory of prudential value? A critical discussion of the desire theories

The main purpose of this chapter is to find out whether the desire theory (in any of its versions) is a justified (or well-founded) theory of prudential value. The focus will be on the satisfaction interpretation, rather than on the object interpretation, of the theory[1]. And since we have assumed that it is not possible to give a satisfaction interpretation of a desire theory unless it is an actual desire theory[2], this implies that we will focus on actual, rather than idealized, desire theories[3].

---

[1]The main reason for this is that the former theory is a substantive evaluative theory (it makes substantive claims about value-for), while the latter theory is a formal theory (rather than making substantive claims about what is good and bad for us, it tells us how to determine what is good and bad for a person) (cf. pp 26-29).

[2]That is, on my view, all idealized desire theories presuppose the object interpretation. Or alternatively put, if we accept the satisfaction interpretation, we also have to accept (D1)(cf. e.g., p 35), but if we accept the object interpretation, we need not accept the counterpart to (D1), viz. the idea that if X is nonderivatively good for P, then it is desired by P (this idea is a part of (OD1) on p 40). The main reason why idealized desire theories cannot be given satisfaction interpretations is that it would be implausible to attribute nonderivative value-for-P to queer entities like the conjunction "P would have intrinsically desired that X obtains if he were rational (informed, or the like) *and* X obtains".

[3]However, the fact that the idealized desire theory (even though it is not a substantive theory of prudential value) constitutes a threat against all actual desire theories means that we can not ignore it altogether. We will take a closer look at the idealized desire theory on pp 275-278.

At this point, we may also add that as long as we are concerned with questions of plausibility, and as long as it is *actual* desire theories that we are interested in, it doesn't really seem to matter which of the two interpretations that we have in mind. For example, most arguments for or against an actual desire theory do not presuppose a certain interpretation, they "support" or "hit" both interpretations of the theory equally well. Now, there are arguments that are for or against one of the interpretations only (e.g., the idea that desire is not prior to value, which is only directed against the object interpretation of the actual desire theory), but these arguments all seem to be "metaethical" or "value theoretical" (or perhaps "justificatory"), rather than evaluative (or normative), in character. On the issue of

So, what we primarily want to know is whether there are good reasons for regarding any version of the actual desire theory as valid (or true). Or more specifically, what we want to find out is this:

(i) What possible version of the actual desire theory is the most plausible theory of prudential value, the unrestricted version or some kind of modified (e.g., restricted) version?

(ii) Is this (most plausible) version of the theory a plausible theory of prudential value? This question is intimately connected to, but not identical with, the question of whether (D1) ("the thesis of actuality") can be justified[4].

We will deal with (i) first, and save (ii) until later. When we deal with (i), we will (so to speak) put (ii) in brackets, or more specifically, we will assume that there is such a thing as a plausible version of the actual desire theory, i.e. that (D1) is valid, that nothing but actual desire-fulfilment can have nonderivative value for a person. It might not be possible to avoid (ii) entirely when discussing (i), however. Or to be more precise; it seems that we need to introduce the "arguments" for (D1) at a pretty early stage. The reason for this is that it might not be possible to conduct the discussion of (i) (what version of the theory that is most plausible) independently of these "arguments". In particular, there are certain arguments for the unrestricted version of the theory (and against alternative versions) that do not just presuppose that (D1) is correct, but that also seem to presuppose that it can be backed up by certain specific "arguments"[5].

---

which interpretation that is to be preferred, cf. the debate in Rabinowicz and Österberg (1996).

[4]Here, it should be kept in mind that these questions might both be asked from different normative perspectives, and that the plausibility of various answers might be "strongly influenced by the point of view of the question" (cf. p 24 and appendix B). For example, it is possible that some version of the desire theory is the most plausible theory of prudential value from the first-person-perspective but not from the third-person-perspective (or in an interpersonal normative context), and it is possible that the version of the desire theory that is most plausible from the first-person-perspective differs from the version that is most plausible from the third-person-perspective (or in an interpersonal normative context). I don't think it will be necessary to make any use of this "insight", however (cf. e.g., p 25).

[5]There is also another reason for presenting the arguments for (D1) at an early stage, viz. because it helps the reader to understand what is attractive about the actual desire theory.

So, why don't we begin with the question of whether (D1) is plausible: why don't we also check out what counts *against* (D1) at this stage? After all, if (D1)

So, what version of the actual desire theory is most plausible? In order to find an answer to this question, we will first take a critical look at the only version of the actual desire theory with which we are (so far) familiar, viz. the unrestricted desire theory[6]. So, is the unrestricted desire theory a plausible version of the desire theory?

5.1. A critical discussion of the unrestricted desire theory

To recapitulate, these are the claims of the unrestricted desire theory (as it has been formulated above):

(D1) The only thing that can be nonderivatively good for a person is to have his actual intrinsic now-for-now desires fulfilled (and so on).

(UD2) The thesis of Unrestrictedness: There are no (intrinsic, now-for-now) desires that it is not nonderivatively good for a person to have fulfilled (and so on).

(UD3) To the extent that it is possible to determine just how strong our desires and aversions are: The nonderivative value (positive or negative) that it has for a person P to have the desire that X (or aversion to X) fulfilled is proportional to the utility of X (for P). That

---

would turn out to be implausible, we would know that there is no plausible version of the actual desire theory, and we would not have to bother with trying to find out what version that is most plausible. Well, let us put it this way: There are people who accept (D1), and my answer to (i) is written with these people in mind. Moreover, there is (as we will see) more to the question "Is the most plausible version of the actual desire theory plausible?" than whether or not (D1) is plausible, and this "extra stuff" may well help us to decide which theory is the best alternative to the actual desire theory, an idealized desire theory or some "objective" theory (this claim might seem unintelligible at this point, but it will hopefully start making sense in due time).

[6]We will not introduce any other (alternative) versions of the actual desire theory until we have seen how the unrestricted version can be (and has been) criticized. The reason for this is that these alternative versions are best understood as modifications of the unrestricted theory, modifications that have (roughly speaking) been made in order to deal with certain objections that have been directed against the unrestricted theory (cf. p 154 above). Why "roughly speaking"? Well, every argument against the unrestricted desire theory can also (at least if we take (D1) for granted) be regarded as an argument for some modified version of the theory, and we should (for this reason) not get the impression that the objections directed against the unrestricted theory are (so to speak) independent of the modified theories.

is, to the extent that our desires and aversions are comparable with respect to strength: It is nonderivatively better for a person to have his intrinsic desire that X (or aversion to X) fulfilled than to have his intrinsic desire that Y (or aversion to Y) fulfilled if and only if he intrinsically prefers X to Y. This is the "intensity-orientation" of the theory.

(UD4) To the extent that it is possible to determine how well off a certain person is: The value that a certain life has for the person who is living it is a function of how much desire-fulfilment and how much aversion-fulfilment that this life "contains", i.e. of how many of his (intrinsic, now-for-now) desires and aversions that are fulfilled, and of how strong these (fulfilled) desires and aversions are. The more desire-fulfilment and the less aversion-fulfilment a life contains, the better this life is for the person who lives it.

So, in order to find out (i) whether the unrestricted theory is a plausible theory of prudential value, we have to find out whether (D1)-(UD4) are well-founded claims, and in order to find out (ii) whether the unrestricted theory is a plausible version of the actual desire theory, we have to assume that (D1) is valid, and then ask whether (UD2)-(UD4) should (*given this assumption*) be accepted or rejected[7]. In short, we have to look at the arguments that have (or can) be given for and against these claims, and ask ourselves whether they are good arguments (and when we consider the arguments for and against (UD2)-(UD4), we have to bear it in mind that most of them are based on the assumption that (D1) is valid). Let us start with the pro-arguments.

5.1.1. Arguments for the unrestricted desire theory

Before we look at the pro-arguments themselves, let us make some general remarks on what it would take to justify (or argue successfully for) the unrestricted desire theory.

In order to justify his theory, there are two rather different things that the unrestricted desire theorist needs to do: He needs to show

[7]If we have "subjective justification" in mind, the question is (rather) whether we (I, you, etc.) are, *on the assumption that we are justified in regarding (D1) as valid*, also justified in regarding (UD2)-(UD4) as valid.

that (D1) is justified, and he needs to show that (UD2)-(UD4) is justified, *given* that (D1) is justified. Or more specifically: First, he needs to justify the idea that there is nothing else besides desire-fulfilment and aversion-fulfilment that is nonderivatively good or bad for a person: he needs to give us good reasons for accepting *some* version of the actual desire theory as valid. However, this means that he needs to show that there are good reasons for regarding all the alternative theories (e.g., hedonism) as false. Second, he also needs to show that there are (*on the assumption that (D1) is valid*) good reasons for accepting (UD2)-(UD4) as valid too. However, to show (in this way) that the unrestricted theory is the most plausible version of the actual desire theory involves showing that no other version of the actual desire theory is (on the assumption that (D1) is valid) plausible.

So the question arises: How can unrestricted desire theorist possibly show all this? What type (or types) of arguments would count as good reasons for preferring the actual desire theory to all the alternative theories of prudential value, and what type (or types) of arguments would count as good reasons for preferring (on the assumption that (D1) is valid) the unrestricted theory to all the other versions of the actual desire theory? If we assume that there are (in fact) good reasons for this, what do these reasons have to be like?

Let us first note that it is (probably because of the "subject-oriented nature" of the theory itself) impossible to come up with any subject-oriented arguments for the desire theory, e.g., for (D1)[8]. To see this, consider the following question: "What is it about us (our nature, our constitution) that makes it nonderivatively good for us to have our desires fulfilled?". This is an odd question, and it is rather obvious that it has no satisfactory answer, e.g., it seems odd to say that it is good for us to have our desires fulfilled because we are desiring beings, or because we have a need for desire-fulfilment, or because we desire to have our desires fulfilled. In short, there are no human features that could possibly be of relevance in this context, i.e. there are no facts

[8]However, it might be possible to construct subject-oriented arguments against (D1), and for some alternative (competing) theory, viz. arguments of the following kind: "The fact that human beings have the feature F (e.g., the fact that we are "social beings") suggests that situations of type X (e.g., spending time with one's friends) are nonderivatively good for us, no matter what we want and do not want". We will take a closer look at this type of argument in chapter 7.

about us that could possibly constitute a good reason for the idea that it is nonderivatively good for us to have our intrinsic desires fulfilled. (And it is important to note that it is not just the satisfaction interpretation of the desire theory that can not be justified in a subject-oriented way: There are no subject-oriented arguments for the object interpretation either. Too see this, consider the question "What is it about all persons P that makes it reasonable to believe that X is good for P iff (and because) P desires that X?")[9].

Now, it is not just that it is impossible to give a subject-oriented justification of the desire theory; it is hard to come up with *any* proper arguments, subject-oriented or object-oriented, for the idea that it is sometimes (or always) nonderivatively good for us to have our intrinsic desires fulfilled: What possible answers, if any, are there to questions like "Why does it sometimes have nonderivative value for us to have our intrinsic desires fulfilled?" or "Why is it sometimes nonderivatively good for a person to get what he intrinsically wants?"? It seems that if there are any proper (objective) arguments for (D1) at all, they would have to take the form of counter-arguments against competing theories[10], and if there are any proper arguments for (UD2)-(UD3) at all, they would have to take the form of counter-arguments against modified versions of the desire theory. As I see it, the only plausible alternative to this "negative view" is the idea that there are no proper pro-arguments at all, only "sources of appeal", or "subjective arguments" that can convince certain people that the (unrestricted) desire theory is valid.

Let us now look at the "arguments" that can be given for the unrestricted desire theory. As has already been suggested, it is fruitful to divide these arguments into two groups, viz. (a) arguments for (D1) and against those theories of prudential value which do not accept (D1), i.e. "objective theories" and "idealized desire theories", and (b)

---

[9]This is why we can't say that it is always the case, for *every* type of theory of prudential value, that a complete justification of this theory must be (in part) "subject-oriented". Instead, we should (because of the desire theoretical case) restrict the scope of this justificatory principle to substantive good theories of a more "objective" type (theories which do not contain any essential reference to our intrinsic features): It is only if a theory of prudential value is of this type that it needs to be justified in a "subject-oriented" way (cf. e.g., pp 110-112 above).

[10]Where these counter-arguments may also (at least in part) be of a metaethical nature, e.g., by appealing to the idea that there are no objective prudential values, only prudential evaluations.

arguments which purport to establish that (UD2)-(UD4) are valid, and that all modified desire theories are invalid, given that (D1) is valid. We will start with the first group of arguments.

"Arguments" for (D1) and against other theories

One of the main "reasons" why the desire theory seems intuitively plausible to many people (at least to many "modern westerners") is that it appeals to "the idea of the Sovereign Subject"[11], or what Harsanyi (1982) calls "the Principle of Preference Autonomy", i.e. the idea that "in deciding what is good and bad for a given individual, the ultimate criterion can only be his own wants and his own preferences" (p 55)[12]. Now, it is rather obvious that this appeal to the principle of preference autonomy does not constitute an argument for the desire theory: The desire theory is best regarded as one of several possible interpretations (or specifications) of the principle, and it would be absurd to insist that the claim interpreted can constitute a good reason for accepting a certain interpretation of this claim. Or alternatively, the principle of preference autonomy is merely a rather vague (but catching) reformulation of the desire theory, and as such, it can not be conceived of as an argument for this theory. It still remains to be seen *why* we should accept the idea that "in deciding what is good and bad for a given individual, the ultimate criterion can only be his own wants and his own preferences".

(i) One possible argument for the desire theory (and for the principle of preference autonomy) is this: A plausible conception of well-being must be flexible enough to "allow for the fact that the best lives for different people may contain quite different ingredients" (cf. Scanlon (1993), p 190). Or as Sumner (1996) puts it, a theory of welfare can not be "descriptively adequate" unless it allows for the fact that "rich and rewarding lives come in a variety of forms" (p 18). Now, as we already

[11]This phrase is from an earlier version of Scanlon (1993).
[12]Or as Sumner (1996) puts it, the desire theory is "in tune with the liberal spirit of the modern age, which tend to see human agents as pursuers of autonomously chosen projects. Unlike objective theories, on which the sources of our well-being are dictated by unalterable aspects of our nature, the desire theory offers us the more flattering picture of ourselves as shapers of our own destinies, determiners of our own good" (p 123).

know, the actual desire theory puts very few restrictions on what a life must be like in order to qualify as "finally good" for this person, and it is therefore maximally flexible in this sense. But does this circumstance really constitute a reason for accepting the theory?

This is how one might argue for a "no" answer: "First, it is (as Sumner (1996) points out) true that the desire theory is 'well placed to satisfy our demand for a unified account of the nature of welfare, while allowing for the multiplicity and variety of its sources' (p 122), but this does not mean that the desire theory is the most flexible theory, e.g., it seems that hedonism is just as flexible. And second, even if it happens to be the case that the desire theory is more flexible than all the plausible alternatives, this does not give us any reason for rejecting these theories: The alternative theories are surely flexible enough (this holds for all plausible versions of the 'objective list theory' as well), and by the way, flexibility is not everything."

To this argument, the (unrestricted) desire theorist can respond as follows: "To demonstrate why our theory should be accepted, let us deal with the objective list theory first, and hedonism later. As far as the objective list theory is concerned, it is simply not true that it is flexible enough, not even in its most plausible versions. To see this, consider what it is for a theory of well-being to be flexible enough. This is what Sumner (1996) suggests:

> No descriptively adequate theory of welfare can simply favour /.../ planning over spontaneity, /.../ complexity over simplicity, /.../ civilization over tribal life, /.../ excitement over tranquillity, /.../ risk over safety, /.../ perpetual striving over contentment, /.../ sexuality over celibacy, /.../ companionship over solitude, /.../ religious conviction over atheism, /.../ rationality over emotion, /.../ the intellectual life over the physical, or whatever (p 18).

If we accept the idea that a plausible requirement of flexibility should be this strong[13], we can hardly regard any objective list theory as flexible enough, and all such theories should therefore be rejected. And as far as the hedonistic theory is concerned, it is simply not true that it is as flexible as our theory (e.g., it does not allow for the possibility that an unpleasant life is a good life, regardless of what other features it has).

---

[13]However, this requirement will be questioned in chapter 7, on pp 305-306.

Moreover, there are also (as we have seen) other reasons for rejecting the hedonistic theory. And note that it is not just that the objections made against hedonism (in section 3.2) do not hit the unrestricted desire theory; this theory also offers a possible explanation of why these objections have the force they have[14]. To sum up, both the hedonistic theory and the objective list theory should be rejected".

(ii) On my view, this is the strongest argument for (D1): "Consider an opposing view, such as the idea that it is nonderivatively good for people to feel pleasure, or to be engaged in creative activity. Now, ask yourself: Do you really think it would be good for a person to feel pleasure, or to engage in creative activity, if he did not have the slightest desire to do so? If there is this person who strongly desires to perform routine tasks, and who has an aversion to being creative, do you really think it would be better for this person to be creative? Of course not! Depending on whether we choose to express ourselves in the idiom of the object interpretation or in the idiom of the satisfaction interpretation, we can either conclude that it can not have nonderivative value for a person to engage in creative activity (etc.) unless he desires to do so[15], or that nothing but actual desire-fulfilment can have nonderivative value for a person, i.e. that (D1) is valid". On my view, this is a striking argument, but it remains to be seen whether it is good enough to give us a sufficient reason to accept (D1). We will return to this issue in section 5.3.

Arguments for (UD2)-(UD3) and against modified desire theories

If we assume that (D1) is valid, how can the unrestricted desire theorist possibly show that (UD2)-(UD3) are valid, i.e. that all modified versions of the desire theory are wrong? What possible arguments can our unrestricted desire theorist come up with?

---

[14]Here, it is important to note that many of the objections directed against hedonism in the last chapter is not really compatible with the strong requirement of flexibility that was put forward above. But some of the objections are, and it is these that we should have in mind here.

[15]It is worth noting that from this idea, it is but a short step to (OD1) on p 40, i.e. the stronger claim that it is good for a person to engage in creative activity (etc.) *if and only if* (and *because*) he has a desire to do so.

Well, the only argument I can think of is "the argument from purity". This is how such an argument for (UD2) (the idea that every intrinsic now-for-now desire is relevant) and against all restricted theories, would look like: It is important to keep it in mind that all restricted desire theorists are desire theorists, i.e. that they accept (D1). Now, there is (on the face of it) nothing incoherent about accepting (D1) while rejecting (UD2), but if we go somewhat deeper, we will see that this is a somewhat problematic position. The problem can be formulated as follows (in the idiom of the object interpretation): Suppose we are restricted desire theorists: suppose we accept the idea that if X is good for P, then P desires that X, but we reject the idea that if P desires that X, then X is good for P. But if we reject the idea that desire is sufficient for value, *why* should we accept the idea that desire is necessary for value? Or alternatively put, if we reject the idea that desire is sufficient for goodness, we also have to reject (OD1), i.e. the idea that X is good for P if and only if (and because) P desires that X. But if we reject this idea, what reason do we have for accepting the idea that desire is necessary for goodness? In short, all the modified versions of the desire theory involve some departure from the intuition that made us accept the desire theory in the first place, viz. "the idea of the Sovereign Subject". But if we reject this idea, it is not clear why we should accept (D1). (That is, restrict, and you weaken your position in relation to objective theories and idealized desire theories[16]). It has been assumed that (D1) is valid, however, and if we assume that it is valid "for a reason", it seems that we must also accept (UD2). So, this is the question that the restricted theorist has to face: "If you accept (D1), and if you have a reason for accepting this claim, then you have to be faithful to that reason when you restrict. And if your reason for accepting (D1) is the reason I think it is: why do you restrict at all?"[17].

So, is this a good argument? I think not, and the reason for this is that the purist assumption on which it is based is rather dubious. As has already been pointed out (on p 63), why assume that the most plausible

---

[16]That is, the restricted desire theories are (so to speak) "squeezed" between the unrestricted desire theory and objective theories like quality hedonism. (Just like the self-interest theory can, on Parfit's view, be squeezed to death between consequentialism and the present-aim theory). We will return to this idea in section 5.3, when we look at the arguments against (D1).

[17]This is probably what Tännsjö (1993) has in mind when he claims (on p 78) that there is something *arbitrary* about all restrictions of the class of relevant desire.

theory of prudential value is a simple theory? After all, we want to find the theory that best fit our semantic, evaluative, and normative intuitions[18], and it is not likely that this is a simple theory. In particular, it is unlikely that the most plausible version of the desire theory is a simple theory. As Scanlon (1993) suggests (on p 187), if we want a desire theory that is in line with "the ordinary meaning of the phrase 'quality of a person's life'", and that (at the same time) "preserves the idea that any improvement in a person's well-being has positive ethical value", then it seems that we simply have to reject the unrestricted version of the theory.

This ends the list of arguments for the unrestricted actual desire theory. Let us now look at how this theory can be criticized.

5.1.2. Arguments against the unrestricted desire theory

Before we take a closer look at (some of) the arguments that can be directed against the theory, let us first make a few general remarks about what these arguments are (in fact) like, and what they have to be like in order to be successful. What would it take to refute the unrestricted desire theory?

Arguing against the unrestricted desire theory: A few general remarks

Let us first point out that just as the unrestricted theory can not be justified in a subject-oriented way, neither can it be refuted (or criticized) in this way (there is a possible exception to this rule, however; cf. note 8 above). This suggests that most (perhaps all) arguments against the unrestricted desire theory can be regarded as attempts to show that the theory has unacceptable evaluative and/or normative implications, i.e. that the theory has implications which are inconsistent with certain valid (or widely accepted) evaluative and/or normative judgements[19]. As a

---

[18]Cf. Sumner's (1996) idea (e.g., on pp 8-10) that a theory of welfare must be both normatively and descriptively adequate.

[19]So, what about the idea that a theory of well-being should be consistent with our "semantic judgements", or more specifically, that it must not deviate too much from the ordinary meaning of the phrase 'quality of a person's life'"? Is this a third

rule, the judgements against which the theory is tested are more specific than the theory itself. They are rarely "fully particular" (i.e. of the form "the particular fact X is nonderivatively good for a particular person P"), however; at times, they are even "fully general", e.g., as in the following argument: "The desire theory implies that 'what you don't know may well hurt you', but this is not so, and the theory must therefore be rejected".

At this point, it is important to keep in mind what our present aim is, viz. to determine whether the unrestricted desire theory is a plausible version of the desire theory (*qua* theory of prudential value). This fact, i.e. the fact that the theory we want to put to the test is a theory of prudential value, puts certain restrictions on what a normative or evaluative intuition must be like in order to qualify as a "possible falsifier" of the theory. So, we have to ask ourselves what kinds of evaluative and normative judgements that are of relevance in this context. What kinds of judgements should be included in the class of possible falsifiers (the class of judgements against which a theory of prudential value might, and ought to be, tested), and what kinds of judgements should be excluded from this class? This is how I think this question should be answered:

The intuitions that are most relevant in this context are (of course) certain kinds of evaluative intuitions, viz. non-comparative intuitions about what desires and aversions it is nonderivatively good and bad for the desiring subject to have fulfilled, and comparative intuitions about which of two relevant desires that it is better for the desiring subject to have fulfilled.

As I see it, there two kinds of normative intuitions that are of special interest in this context, viz. the following ones:

(i) Intuitions about what kinds of desires that it is (on the desire-fulfilment version of the self-interest theory of rationality) rational to try to fulfil[20], or more precisely, intuitions about what version of

---

kind of judgement with which a theory of well-being must be consistent? I think not. On my view, it is impossible to distinguish our evaluative intuitions from our semantic intuitions. This could explain why Sumner (1996) includes both in his criterion of "descriptive adequacy", the idea that a theory of welfare must "fit /.../ our ordinary experience of welfare and our ordinary judgements concerning it" (p 8).

[20]Here, I ignore the possibility that trying to achieve what is good for one might make one worse off than one would otherwise be.

"preference-egoism" that is the most plausible theory of rational choice.

(ii) Intuitions about what kinds of desires that it is morally right (from a preference-utilitarian point of view) to try to fulfil[21], or more precisely, intuitions about what version of "preference-utilitarianism" that is the most plausible moral theory[22].

We must not assume that all intuitions of type (ii) are of relevance in this context, however. As an illustration of this, consider the case of sadistic desire. It seems reasonable to assume that on the most plausible version of preference-utilitarianism, such desires should not be taken into account. This does not imply that it is never good for us to have our sadistic desires fulfilled, however, and neither does it imply that it is never rational to try to fulfil a sadistic desire. (The last wishes of the dead is another example).

The following kinds of intuitions are most certainly irrelevant in this context, i.e. they must not be regarded as possible falsifiers of a theory of prudential value:

(i) Intuitions about what kinds of desires and aversions that it is (on the instrumental theory of rationality) rational for the agent to "act upon" (e.g., try to fulfil), or more precisely, intuitions about what version of the instrumental theory that is the most plausible theory of rational choice. The fact that intuitions of this kind are irrelevant has certain implications, e.g., the following one: Even though it may well be the case that "acting rationally means acting consistently on beliefs and desires that are not only consistent, but also rational" (Elster (1983), p 15), this does not allow us to conclude that it is never good for us to have our irrational desires fulfilled.

(ii) Intuitions about what kinds of desires and aversions that should be taken into account by a utility-oriented theory of social justice, i.e. intuitions about what kind of utility that should be justly distributed between the members of a society (assuming that it is utility, viewed as preference-fulfilment, and not something else, that should be justly distributed).

---

[21]Here, I ignore the possibility that trying to achieve what is good-period might actually result in less good than some of the alternative courses of action. Cf. note 20.

[22]These are not the only relevant normative intuitions, however, e.g., intuitions about what one should do in order to benefit one's children or friends must also be regarded as relevant. Cf. also appendix B.

(iii) Intuitions about what version of "the social choice theory" that is the most plausible theory of collective rational choice, i.e. intuitions about what kinds of individual preferences that should serve as inputs to a collective choice if this choice is to be regarded as rational (assuming that a collective decision can not be rational unless it reflects the actual preferences of the members, that is). To see what is implied by the fact that intuitions of this kind are irrelevant, consider the following case: Even though Elster (1983) may well be correct in claiming that "[f]or the purpose of social choice theory, we should not take wants as given, but inquire into their rationality or autonomy" (p 140), this does not allow us to conclude that it is never nonderivatively good for us to have our adaptive intrinsic desires fulfilled.

To sum up, the important thing to notice here is that there are several evaluative and normative contexts in which desires and preferences might be of interest, and that preferences that are of relevance in one of these areas may well be irrelevant in the other areas. If we are not aware of this fact, we may take intuitions we have about one area and try to use them in another area, something which will most probably result in a terrible muddle.

The arguments against the unrestricted desire theory: A brief overview

There are at least two different ways in which the objections against the unrestricted actual desire theory can be classified. (1) First, they can be classified on basis of what claim or claims that is the primary target, i.e. on basis of what exactly that they purport to refute (or disprove). If we classify the objections in this way, we end up with the following categories: Arguments which are primarily directed against (D1), arguments which are primarily directed against (UD2), arguments which are primarily directed against (UD3), and arguments which are primarily directed against (UD4)[23]. (2) However, the objections to the unrestricted desire theory (particularly the objections to (UD2) and (UD3)) can also be divided into the following three categories: Object-oriented (or content-oriented) objections, rationality-oriented objec-

[23]However, it is worth pointing out that many of these "counter-arguments" are (as we will soon see) really "counter-claims" rather than proper counter-arguments.

tions, and objections that are neither object-oriented nor rationality-oriented. So, let us now take a closer look at these categories of objections.

(1) If we classify the objections on basis of what claim or claims that they purport to refute, they will fall into the four categories mentioned above:

(i) Arguments against (D1), i.e. arguments that purport to show that there are other things besides actual desire-fulfilment that can be non-derivatively good for a person. All these arguments are (at the same time) arguments for some alternative theory of prudential value, either for some objective theory, or (if we remain within the framework of the desire theory) for some idealized version of the desire theory. However, since the present aim is to find out which version of the actual desire theory that is most plausible (given that (D1) is true), we will not deal with these arguments until later.

(ii) Arguments against (UD2) (the thesis of unrestrictedness), i.e. arguments that purport to show that there are actual intrinsic desires which it is not good for the desiring subject to have fulfilled. Most of these arguments are based on the assumption that (D1) is true, and some of them also takes (UD3) for granted. It is important to observe that *if* (D1) is taken for granted, then every argument against (UD2) is (at the same time) an argument for some restricted version of the desire theory. Or alternatively put, if (D1) is not questioned, then every argument against (UD2) can be viewed as being delivered from the standpoint of some kind of restricted desire theory: it can be viewed as part of a "discussion" that is (so to speak) internal to the desire theory[24].

(iii) Arguments against (UD3), e.g., arguments that purport to show that it is not always the case that the stronger a (relevant) desire is, the better it is for the desiring subject to have it fulfilled. It seems that all arguments of this kind take (D1) for granted[25], and furthermore, that they are not really directed against the intensity-orientation itself, but

---

[24]So, what about arguments that purport to show that what people actually want may be "grossly against their interests" (cf. Parfit (1984), p 127)? Should these arguments be classified as arguments against (D1) or as arguments against (UD2)? Well, on my view, all such arguments hit (if successful) (UD2), but I also suspect that most (or all) of them hit (D1).

[25]Whether or not (UD2) is taken for granted here does not matter.

the "actual intensity-orientation" of the theory, or the conjunction of (UD3) and (D1). If it is true that (D1) is always taken for granted in this way, then every argument against (UD3) is (at the same time) an argument for some non-intensity version of the actual desire theory, i.e. for some specific claim concerning relative weights[26].

(iv) Arguments directed against (UD4), i.e. arguments that purport to show that we should (roughly speaking) not regard a person's overall level of well-being as a function of how much desire-fulfilment and aversion-fulfilment there is in his life. Now, the counter-arguments I have in mind here are evaluative (or normative) rather than methodological[27], i.e. they purport to show that even if (UD4) were methodologically sound (e.g., even if it were possible to individuate a person's desires in a proper way, calculate the relevant sums, and so on), it would not be valid anyway. It is worth noting that these arguments are *not* (at the same time) arguments for any particular alternative view, and they do not "in themselves" suggest any alternative (evaluative) claim that could replace (UD4). However, there are certain views that have been put forward in response to the criticism of (UD4), but most of these views can (as we will see on pp 217-219) hardly be considered as possible replacements of this claim.

(2) Let us now take a closer look at the other possible way in which the objections to the unrestricted actual desire theory can be classified. To classify the counter-arguments in this way is (as we have already seen) to divide them into the following three categories:

---

[26]It is worth noting that for every possible objection against (UD3), there is a corresponding objection against (UD2) (and vice versa), and that objections of the latter kind are always stronger than objections of the former kind. For example, the objections against (UD2) that purport to show that it is never good for us to have our irrational desires fulfilled are stronger than the corresponding objections against (UD3), viz. the objections that purport to show that it might be better for a person to have a rational desire fulfilled than to have an irrational desire fulfilled, even if the latter desire is stronger. As we will see below, this circumstance has certain implications for in what order the objections against the unrestricted theory are best presented.

At this point, it is also worth noting that the three types of objections (i)-(iii) correspond to the three possible ways in which the unrestricted actual desire theory can be modified, viz. (i) idealizations, (ii) restrictions (or eliminations), and (iii) modifications of the intensity-orientation, i.e. modifications which result in "non-intensity versions" of the actual desire theory.

[27]There are (as we already know) strong methodological arguments against (UD4), arguments that have been "sketched" on pp 155-159 above. There will be no further elaboration of these arguments, however.

(a) Object-oriented objections, i.e. objections that purport to show that desires with certain kinds of objects (or propositional contents) should be regarded as more relevant than desires with other kinds of objects (contents). Objections of this kind are either directed against (UD2), the thesis of unrestrictedness, or against (UD3), the idea of proportionality (which includes the intensity-orientation). In the former case, it is claimed that desires with the "wrong kind" of content (e.g., desires that are not about the life of the desiring subject) should not count at all, and in the latter case, it is claimed that desires with the "right kind" of content should be regarded as more relevant than desires with the "wrong kind" of content.

(b) Rationality-oriented objections, i.e. objections that purport to show that it is (roughly speaking) better for us to have our rational desires fulfilled than to have our irrational desires fulfilled. Again, objections of this kind can be directed both against (UD2) (they may purport to show that all irrational desires should be regarded as irrelevant) and against (UD3) (they may "merely" purport to show that rational desires should be regarded as more relevant than irrational desires). However, it is important to note that there are also rationality-oriented arguments against (D1), viz. when it is claimed that if P *would* desire that X under ideal circumstances (e.g., if he were rational or fully informed), then it is good for P that X obtains, even if P does not actually desire that X[28].

(c) Objections that are neither content-oriented nor rationality-oriented, e.g., objections which appeal to the idea that we should give more weight to those of our desires that are sanctioned by higher-order desires, or to those of our desires that we know that we have. Objections of this third kind are (again) either directed at (UD2) (e.g., when it is claimed that all desires that are not sanctioned by higher-order desires should be regarded as irrelevant, or that it can not be good for a person to have a desire fulfilled unless he is aware of the occurrence of the object), or against (UD3) (e.g., when it is claimed that desires which are known to the desiring subject should be regarded as

---

[28]It should be noted that a certain overlap between (a) and (b) is possible, that an objection can be object-oriented and rationality-oriented at the same time, but only if there are desires that are rational or irrational in virtue of their content. We might also add that if this is so, then there is a sense in which objections against (D1) might be object-oriented.

more relevant than desires that are unknown to the subject).

At this point, it should be rather obvious that the two classifications cross each other, e.g., objections of type (b) can be of type (i), but they can also be of type (ii) or (iii). The resulting two-dimensional classification of the objections can be illustrated as follows:

	against (D1)	against (UD2)	against (UD3)
object-oriented	? (cf. note 28)	X	X
ratio-oriented	X	X	X
other kinds	-	X	X

I have chosen to present the objections "from top to bottom" rather than "from left to right". My reason for this is twofold. First, claims like (UD2) and (UD3) are rarely criticized in isolation, and it is (for this reason) rather problematic to classify the objections on basis of what claim they purport to refute. And second, every possible objection against (UD2) is a stronger version of some possible objection against (UD3), and every possible objection of the latter kind is a weaker version of some possible objection of the former kind (cf. note 26), and it seems (for this reason) reasonable to put the criticism against (UD2) and the criticism against (UD3) together (they are, after all, based on the same vague intuitions). And if we, on top of this, remind ourselves that most of these objections are based on the assumption that (D1) is true, we can see that it is not a very good idea to let the presentation of the arguments follow the first classification.

So, this is how the arguments against the unrestricted actual desire theory will be presented: First, we will look at some object-oriented objections, then at some rationality-oriented objections, and finally at some objections that are neither object-oriented nor rationality-oriented. When we do this, we will, to the extent that it is possible, distinguish (within each group) between objections against (UD2) and objections against (UD3). (The objections against (D1) will be saved until later). The last type of counter-argument that we will look at are the evaluative arguments against (UD4), arguments which cannot be classified as object-oriented, rationality-oriented, or the like. So, let us now look at the counter-arguments themselves. We will start with the first group of arguments, viz. the content-oriented (or object-oriented)

objections.

Object-oriented objections to the unrestricted actual desire theory

(i) The first argument is from Parfit (1984), and it is primarily directed against (UD2). Parfit writes:

> Suppose that I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Later, unknown to me, this stranger is cured. On the Unrestricted Desire-Fulfilment Theory, this event is good for me, and makes my life go better. This is not plausible. We should reject this theory (p 494).

Now, as it stands, this argument does not necessarily hit the thesis of unrestrictedness (as we have formulated it). Or more specifically, it is possible that the argument works for some other reason, e.g., because the benevolent desire in the example is prospective, or because the subject does not know that it has been fulfilled. However, it only takes a slight modification of the example to achieve the desired effect: "A little later, while I still have the (intrinsic) desire, the stranger is cured. This does not have nonderivative value for me (not even if I happen to find out about it), and we should therefore reject the unrestricted desire theory". This is a strong objection to the theory.

(ii) The first object-oriented argument suggests that benevolent desires of a certain type are irrelevant. Isn't it even more certain that certain malevolent desires should be regarded as irrelevant[29]? Suppose that Alice has an intrinsic now-for-now desire that Bert suffers. Is it really nonderivatively good for Alice to have this desire fulfilled? I think not. That is, it is not just that Alice's desire is irrelevant in the "social", or "interpersonal" context, e.g., that it can safely be ignored by a social choice theorist; it should also (and this is a different thing) be regarded as irrelevant in the present "intrapersonal" context. But notice that the

[29]This is but a special version of a more general idea, viz. that all immoral desires (desires which do not survive moral criticism) should be regarded as irrelevant. This is not a very plausible idea, however. For example, if I have an intrinsic desire to perform an action that happens to be morally wrong, it is far from certain that it is not good for me to perform the action.

reason why Alice's desire should be regarded as irrelevant may not be the same in the two cases. For example, the reason why it should not count in the social context may be that it is "anti-social" (cf. Harsanyi (1982), who claims that "all clearly antisocial preferences, such as sadism, envy, resentment, and malice" must be "altogether excluded from our social-utility function" (p 56)), while the reason why it should not count in the present context might have more to do with the fact that it is other-regarding[30].

(iii) Suppose that Bert has the following two intrinsic now-for-now desires: A global desire to work as a doctor, and a local desire that Alice (to whom he is just talking) will take a fancy to him. Suppose also that the local desire is stronger. Isn't it, in this case, really better for Bert to have the weaker (global) desire fulfilled? Well, I think not, not if we have the relevant notion of strength in mind, i.e. place in a preference ordering. First, if we have the relevant conception of strength in mind, it is highly unlikely that our local desires will be stronger than our global ones, and second, even if they are (at times) stronger, this is no reason to reject the thesis of intensity. If Bert's desire to be liked by Alice is (in the relevant sense) stronger than his desire to work as a doctor, we should accept the idea that it is better for Bert that Alice likes him. (Better for Bert-at-that-time, that is. What is better for Bert-the-next-day is another matter).

The Success Theory, the experience-oriented Success Theory, and the global Success Theory are examples of modified desire theories which, in different ways, respond to objections of the object-oriented type. The different object-oriented versions of the desire theory (which need not necessarily be restricted, they can also be of a "non-intensity" kind) will be discussed in section 5.2.1.

---

[30]The reason why the desire in the example is truly other-regarding is that it does not contain any reference at all to Alice. That is, if Alice would (instead) have desired that *she* makes Bert suffer, that he suffers as a result of *her* actions, the desire would not have been other-regarding.

The idea that all other-regarding desires should be conceived of as irrelevant (in this context) has several attractive features, e.g., it would also explain why it is not good for Parfit to have the benevolent desire in (i) fulfilled, and it would imply that we do not have to bother with questions like "Is it better for me to have my benevolent desires fulfilled than to have my malevolent desires fulfilled?". But is the idea valid? This will be investigated in section 5.2.1 below.

Rationality-oriented objections to the unrestricted actual desire theory

If we (for the time being) ignore the objections to (D1), we know that the rationality-oriented objections to the unrestricted desire theory are of two fundamental kinds, viz. objections that purport to show that it has no value whatsoever for a person to have his irrational desires fulfilled (i.e. arguments against (UD2)), and objections that purport to show that it is (roughly speaking) better for a person to have his rational desires fulfilled than to have his irrational desires fulfilled (i.e. arguments against (UD3)). Or alternatively put, the basic idea is that those of our desires which do not survive rational criticism must either be regarded as totally irrelevant (their fulfilment does not make us better off at all) or less relevant than those of our desires that survive such criticism.

In this (critical) section, it is not really necessary to know what exactly that the rationality (or irrationality) of a desire consists in[31]. None of the objections below presupposes any particular conception of rational desire; they all focus on desires that are irrational on several conceptions, e.g., desires which are based on beliefs that are both false and irrational, which the subject would not have "if he knew the relevant facts, was thinking clearly, and was free from distorting influences", and which have (on top of this) "the wrong kind of causal history". So, here are the objections:

(i) Suppose that John has an intrinsic desire to spend time with Paul, and that this desire is derived from the intrinsic desire to spend time with his friends and the false (and irrational) belief that Paul is his friend. Is it really nonderivatively good for John to spend time with Paul and have his desire fulfilled? Is it as good as it would be if his belief were true, and Paul were (in fact) his friend? I think not.

---

[31]However, it is worth noting already at this point that there are at least five different conceptions of rational desire, conceptions which give rise to five different "senses" in which a desire might be rational (irrational), viz. (i) Hume's theory, (ii) the informed theory, (iii) the deliberative theory, (iv) the genetic theory, and (v) the intrinsic theory. We will take a closer look at these conceptions in section 5.2.2, where we will also discuss what notion of rational desire that is of most relevance in this context, i.e. what conception of rational desire that makes a rationality-oriented modified desire theory most plausible.

(ii) Suppose that Bert has a strong intrinsic (now-for-now) desire to work for the CIA (MI6, SÄPO, or the like). Suppose also that this desire has been produced by watching movies of a certain kind, and that there is no way in which it could have been produced "naturally", i.e. "by experience with the actual situations which the desire is for". Does it, in this case, really have nonderivative value for Bert to work for the CIA? I think not.

(iii) Suppose that Bert's intrinsic desire to be loved and admired by Alice is much stronger than his intrinsic desire to keep his job, and that the reason for this is that the former desire is derived from an abnormally strong desire to be admired and/or "loved" by "everyone". Suppose also that this more fundamental desire was produced by "an early and prolonged deprivation of something wanted". Can the actual desire theorist, in this case, really be certain that it is nonderivatively better for Bert to have the stronger desire fulfilled, i.e. that it is better for Bert that Alice loves and admires him than that he keeps his job? (Notice that this is an argument against (UD3) rather than against (UD2)).

(iv) Suppose that Alice intrinsically prefers living alone (as a single) to living a family life. Suppose also that this preference is either adaptive or counteradaptive, i.e. that it is either a result of "adaptive preference formation" (it is caused by "the drive to reduce the tension or frustration that one feels in having wants that one cannot possibly satisfy"), or a result of "counteradaptive preference formation" (Alice lives a family life and believes that "the grass is always greener on the other side of the fence"). Is it, in both these cases, nonderivatively better for Alice to live as a single? I'm not quite sure that it is, especially not if we have the relevant notion of strength in mind, i.e. rank in a preference ordering (The argument would surely be more convincing if we had felt intensity in mind; cf. the next argument below).

(v) The following well-known argument is another version of the idea that our adaptive preferences should be regarded as irrelevant. Let us start from the idea that non-adaptive (e.g., autonomous) desires should (*in the context of interpersonal comparisons*, e.g., in the context of social

justice) have more weight than adaptive desires. Both Nussbaum (1990) and Sen (1985) have argued that actual desire is (in this area) "a malleable and unreliable guide to the human good" (Nussbaum (1990), p 213)[32]. Their main reason seems to be that our actual desires are often adaptive desires, and that these desires are of little or no relevance in this context. This criticism is very similar to Elster's (1983) criticism of the "actual preference utilitarian" (who regards individual want satisfaction as "the criterion of justice and social choice"). He writes:

> [W]hy should the choice between feasible options only take account of individual preferences if people tend to adjust their aspirations to their possibilities? For the [preference] utilitarian, there would be no welfare loss if the fox were excluded from consumption of the grapes, since he thought them sour anyway. But of course the cause of his holding them to be sour was his conviction that he would be excluded from consuming them, and then it is difficult to justify the allocation by invoking his preferences (p 109).

This is Nussbaum's (1990) objection to the idea that what is good for people is (in this interpersonal context) "a function of the satisfaction of desires or preferences that people happen, as things are, to have":

> The central difficulty with this proposal is /.../ that desire is a malleable and unreliable guide to the human good, on almost any seriously defensible conception of good. Desires are formed in relation to habits and ways of life. At one extreme, people who have lived in opulence feel dissatisfied when they are deprived of the goods of opulence. At the other extreme, people who have lived in severe deprivation frequently do not feel desire for a different way, or dissatisfaction with their way. Human beings adapt to what they have. In some cases, they come to believe that it is right that things should be so with them; in other cases, they are not even aware of

---

[32]But what is this supposed to mean? Well, if we let P1 and P2 be two desiring subjects, I think there are at least two different interpretations of this claim, viz. (i) "The increment of P1's well-being that 'results from' having a certain actual desire D1 fulfilled may well be larger than the increment of P2's well-being that 'results from' having a certain desire D2 fulfilled, even if D2 is stronger than D1" (This is based on the dubious assumption that the strength of D1 is comparable to the strength of D2); and (ii) "P1 may be better off than P2, even if there is more actual desire-fulfilment in P2's life than in P1's life". This claim seems more like an objection to (UD4), however, or to the desire theory as a whole.

alternatives. Circumstances have confined their imaginations. So if we aim at satisfaction of the desires and preferences that they happen, as things are, to have, our distributions will frequently succeed only in shoring up the status quo (p 213).

This is how Sen (1985) formulates the same argument (notice that he explicitly has the context of interpersonal comparisons in mind):

> Comparative intensities of desire /.../ are influenced by many contingent circumstances that are arbitrary for well-being comparisons. Our reading of what is feasible in our situation and station may be crucial to the intensities of our desires, and may even affect what we dare to desire. Desires reflect compromises with reality, and reality is harsher to some than to others. The hopeless destitute desiring merely to survive, the landless laborer concentrating his efforts on securing the next meal, the round-the-clock domestic servant seeking a few hours of respite, the subjugated housewife struggling for a little individuality, may all have learned to keep their desires in line with their respective predicaments. Their deprivations are gagged and muffled in the interpersonal metric of desire fulfillment. In some lives small mercies have to count big (pp 190-191).

So, is this a good argument? Or more specifically: (a) Is it a good argument in the context in which it "really belongs", i.e. in "the context of interpersonal comparisons"? And (b) if it is, does it also show (as Nussbaum seems to think) that adaptive desires are less relevant than autonomous desires in the present (intrapersonal) context as well? Let us start with the second question:

(b) No, I don't think so. The alleged fact that the actual desire theory is no good in the context of interpersonal comparisons[33] does not allow us to conclude that it is also bad *qua* theory of prudential value (cf. pp 198-200 above).

(a) In order to find out whether the argument is good if viewed in its proper context, we need to proceed with care, viz. for the following reason: The notion of strength which is presupposed in Nussbaum's

---

[33]The alleged fact can be more precisely formulated as follows: "Even if social justice would be a question of how final goods should be distributed, it would certainly not be defensible to accept the actual desire theorist's view on what has final value. And even if it were true that it is utility that we ought to maximize, it would certainly not be utility in *this* sense".

and Sen's formulations of the argument is certainly *not* rank in a cool preference ordering, but felt intensity (or perhaps motivational force); a person's intrinsic preference ordering is not *that* easily affected by his judgements of probability, e.g., by his beliefs about how likely he is to succeed if he tries to fulfil a certain desire. So the question arises: Does the argument work if we have the proper notion of strength in mind? Well, I think it does. How strongly a person wants something in the relevant sense is also (to a certain extent) causally dependent on his beliefs about his own possibilities[34], and this suggests that we should reject the idea that it is actual desire-fulfilment that should be maximized or justly (e.g., equally) distributed.

(vi) Suppose that Bert intrinsically desires to be betrayed, manipulated, slandered, deceived, and to suffer great pain. Is it really nonderivatively good for him to have these desires fulfilled? It can be argued that it is not good for him, for the simple reason that the objects of these desires are not only (in some relevant sense) "in no respect worth desiring", but also "worth avoiding", e.g., because wanting these things is (in some sense) "against human nature"[35].

This ends the section on rationality-oriented objections to the unrestricted desire theory. If any of these objections are valid, the actual desire theorist should accept some kind of rationality-oriented modified desire theory.

Other types of objections to the unrestricted actual desire theory (objections that are neither object-oriented nor rationality-oriented)

As I see it, there are two types of interesting objections which belong to this category. First, there are the objections which appeal to the idea that we should give more weight to those of a person's desires that are

---

[34]The fact that I adjust my aspirations to my perceived possibilities need not have any effect on my preference ordering at all, however. For example, the fact that I believe that it is impossible for me to get rich and therefore lower my expectations does not necessarily make the situation "I am rich" drop on my preference ordering.

[35]It is worth noting that even though this objection is (strictly speaking) object-oriented, I prefer to regard it as rationality-oriented.

sanctioned by his higher-order desires, evaluations, or ideals than to those of his desires that are not sanctioned by his higher-order desires (etc.), e.g., that we should give little or "no weight to the desires that someone wishes that he did not have" (cf. Parfit (1984), p 119)[36]. And second, there are also the "knowledge-oriented" objections, some of which appeal to the idea that we should give less or no weight to those of a person's desires of which he is unaware[37]; while others appeal to the idea that it can not be good for a person to have a desire fulfilled unless he is aware of the occurrence of its object (or of the fact that it is fulfilled)[38].

So, here are a couple of objections of the first kind. The first argument is primarily directed against (UD2):

(i) Suppose that Alice intrinsically desires to smoke a cigarette. Suppose also that she wants to be a non-smoker, or more specifically, that she desires that she does not have the first-order desire to smoke[39]. It might be argued that in a case like this, it is not nonderivatively good for Alice to smoke the cigarette.

(ii) The next objection is directed against (UD3): Suppose that Bert has the following two intrinsic desires, viz. to go trekking in Nepal, and to do some heavy partying in Greece. Suppose also that the two desires are equally strong (in the relevant sense), but that the former desire is to a higher extent sanctioned by his appropriate higher-order desires, evaluations, or ideals than the latter (we can assume that Bert has a strong desire to be a sporty type of person, and we can also, for the sake of the argument, assume that this desire is rational). Isn't it, in this

---

[36]That is, if P wishes that he did not have a certain first-order desire D, then D is *not* sanctioned by P's higher-order desires. But this is only a partial answer to the question of what it is for a first-order desire to be sanctioned (or not sanctioned) by higher-order desires. I will try to give a more complete answer in section 5.2.3.

[37]That we do not always know what we want is a possibility that every plausible conception of desire (functional or not) allows for (cf. section 4.1).

[38]That a person's desires can be fulfilled without him knowing it follows from the notion of fulfilment that was "adopted" in section 4.3.

[39]We can also assume that the second-order desire is rational (Or can we? Cf. note 109 on p 264). It is hardly intrinsic, however; this would be too odd to be realistic. And as far as the relative strengths of the two desires are concerned, I don't think it is necessary to make any assumptions at all: On my view, it has no significance whatsoever which desire is stronger, the second-order desire or the first-order desire.

case, better for Bert to go trekking in Nepal?

Are these good arguments? Well, I am not quite sure about this. It might be suspected that both objections presuppose the wrong notion of desire, viz. the phenomenological conception, as well as the wrong notion of strength, viz. strength as felt intensity. Or alternatively put, if we have the relevant (non-phenomenological) notions of desire and strength in mind, objections of this kind may no longer be valid. Let me explain more in detail: On the relevant notion of desire, someone desires that X if and only if he prefers X to some situation to which he is neutral, and on the relevant notion of strength, someone's desire that X is stronger than his desire that Y if and only if he prefers X to Y (where preference should not be understood in terms of felt intensity). That is, on the relevant (non-phenomenological) notions of desire and strength, both notions are understood in terms of (rankings on) preference orderings. However, the ranking of a certain "first-order situation" (the object of a first-order desire or aversion) in a person's cool preference ordering is necessarily affected by his higher-order desires and ideals in a way that the felt intensities of his first-order desires are not. For example, the fact that someone wants to be a certain type of person is highly likely to have a certain effect on his "first-order preferences". So we may ask ourselves: Isn't it sufficient that our higher-order desires are already "incorporated" into our "first-order preference orderings" in this way? Isn't there something arbitrary about "counting our higher-order desires twice"?

Let us now take a second look at the two objections above. If Alice's intrinsic desire to smoke is a desire in the relevant sense, then she actually prefers smoking a cigarette to some situation to which she is neutral (and she prefers smoking to non-smoking), in spite of the fact that she wishes that she does not have the first-order desire to smoke. Isn't it, in this case, good for Alice to have a smoke? Or consider the case of Bert. If we have the relevant notion of strength in mind, Bert is indifferent between going trekking in Nepal and partying heavily in Greece, in spite of the fact that he strongly prefers to be an outdoor type of person to being a "party animal". Can we, in this case, really conclude that it is nonderivatively better for Bert to go trekking in Nepal? Is it really plausible to count Bert's ideal twice?

Let us now look at an objection of the second kind. The argument is

directed against (UD2):

(iii) Suppose that Bert has a strong intrinsic desire to be admired by Alice, but that he does not know (or believe) that he has the desire. Suppose also that Alice does (in fact) regard Bert with esteem, but that he has no awareness whatsoever of this fact. Is it, in this case, really nonderivatively good for Bert to have the desire fulfilled? I think not.

But why is this? For example, is it because he does not know (believe) that he has the desire, or is it because he does not know (believe) that it has been fulfilled, i.e. that the desired situation obtains? Well, this remains to be seen.

We will return to these issues in section 5.2.3. In that section, we will also take a look at Kekes' version of the desire theory, which is a good example of a theory that responds to objections of the first type (regarded as valid).

Evaluative objections to (UD4)

As we already know, (UD4) is a claim about how a person's overall level of well-being should be determined. How well off a person is (on the whole) is, on this view, a function of how many of his intrinsic now-for-now desires and aversions that are fulfilled, and of how strong these desires and aversions are: The more desire-fulfilment there is in his life, the better-for-him, and the more aversion-fulfilment there is in his life, the worse-for-him. We have also seen (on p 158 above) that there are at least two possible specifications of this idea, viz. "the difference thesis" and "the ratio thesis".

Now, in order to isolate the evaluative criticism of (UD4) from the methodological criticism of the same claim, and from the objections to (D1)-(UD3), it is necessary to make certain assumptions: (i) First, we have to assume that the claim is methodologically sound, i.e. that it is possible to formulate it in a meaningful and intelligible way (which means that we have to assume that it is possible to get hold of the two sums Σ (df) and Σ (af) on p 156 above, that it is meaningful to perform arithmetic operations on these sums, and so forth). (ii) We also have to

assume that (D1)-(UD3) are valid claims[40].

Here are a few arguments which are (or seem to be) of this type:

(i) If all the necessary assumptions are made, (UD4) implies that a person can improve the quality of his life simply by wanting (intrinsically) what happens to happen, or by eliminating intrinsic aversions to what is actually the case[41]. This is counter-intuitive, however: Bert can not improve the prudential value of his life by intrinsically desiring that the sun rises in the morning or that it sets at night.

As I see it, this is a good argument against the unrestricted desire theory, but we may ask what component of the theory that it hits. It is fruitful to view (UD4) as a combination of two separate ideas, viz. (UD2), the thesis of unrestrictedness, and (UD4'), the version of evaluative atomism which conceives of a person's overall level of well-being as a function of how much *good* desire-fulfilment and how much *bad* aversion-fulfilment there is in his life, i.e. as a function of how many of his *relevant* desires and aversions that are fulfilled, and of how strong these desires and aversions are (cf. p 39). Now, once we become aware of this, I think it is pretty clear that the objection above hits (UD2) rather than (UD4'). All the argument shows is that we should give no weight to the desire that the sun rises in the morning: it does not give us any reason to assume that the value of a person's existence as a whole is not a function of the values of its parts. That is, *if* (D1) and (UD3') are valid claims, and *if* we make the appropriate restrictions (this is to ensure that we are dealing with valuable desire-fulfilments), then there is no reason to believe that (UD4') is false.

(ii) On my view, the same thing holds for the next argument against

---

[40]So, what the evaluative objections to (UD4) should purport to show is that it is an implausible evaluative claim, *even if* it is intelligible, and *even if* (D1)-(UD3) are valid. This makes it rather doubtful whether there are any plausible arguments of this kind, i.e. arguments which are directed exclusively against (UD4) (or against (H6), for that matter!). As we will soon see, the primary target of the arguments below need not be (and might not be) (UD4).

[41]It is important to notice that (UD4) does *not* imply that we can improve our lives by not wanting what we are unlikely to get, e.g., by eliminating desires which are difficult to satisfy, or whose satisfaction is uncertain. The reason for this is simple: A plausible desire theory does not claim that it is bad for a person not to have his desires fulfilled.

(UD4). The argument is from Parfit (1984):

> Consider this example. Knowing that you accept a Summative theory, I tell you that I am about to make your life go better. I shall inject you with an addictive drug. From now on, you will wake each morning with an extremely strong desire to have another injection of this drug. Having this desire will be in itself neither pleasant nor painful, but if the desire is not fulfilled within an hour it will then become very painful. This is no cause for concern, since I shall give you ample supplies of the drug. Every morning, you will be able at once to fulfil this desire. The injection, and its after-effects, would also be neither pleasant nor painful. You will spend the rest of your days as you do now.
>
> /.../ We can plausibly suppose that you would not welcome my proposal. /.../ But it is likely that your initial desire not to become addicted, and your later regrets that you did, would not be as strong as the desires you have each morning for another injection. Given the facts as I describe them, your reason to prefer not to become addicted would not be very strong. /.../
>
> On the Summative Theories, if I make you an addict, I will be increasing the sum-total of your desire-fulfilment. /.../ On the Summative Theories, by making you an addict, I will be benefiting you - making your life go better.
>
> This conclusion is not plausible. Having these desires, and having them fulfilled, are neither pleasant nor painful. We need not be Hedonists to believe, more plausibly, that it is in no way better for you to have and to fulfil this series of strong desires (p 497).

Now, as I see it, this is a good argument, i.e. it seems pretty obvious that living as a successful addict is not better for the person than having a drug-free life. But in this case too, it seems that the argument does not really hit (UD4'), as Parfit seems to believe, but (UD2). As Parfit himself points out, the "Summative Theorist" can always respond to the objection by claiming that there is some feature of the addictive desires which makes them irrelevant, e.g., by suggesting that these desires "can be ignored because they are desires that you would prefer not to have" (cf. ibid., pp 497-98). Parfit rejects this idea, however, but his

argument for this (on p 498) is defective[42], he does not show that it is good for the person to have his addictive desires fulfilled, and it is (on top of this) likely that he does not have the relevant notions of desire and strength in mind. Moreover, Parfit's case against (UD4') gets even weaker if we see how Parfit himself tries to improve it. His own "solution" to the "addiction objection" is to appeal to a so-called Global version of the desire theory, i.e. to regard people's local desires as irrelevant. But even if we (implausibly) assume that we can exclude all local desires in this way, this is hardly anything that counts against (UD4'). And the problem of aggregation still remains, viz. in the following form: If we know how many of a person's global desires and aversions that are fulfilled, and how strong these (fulfilled) desires and aversions are, how do we use this information in order to determine how well off he is? If we reject (UD4'), it is doubtful whether there is anything with which it can be replaced. Or is there?

If we assume that there is a better alternative to (UD4'), what would such an alternative notion of aggregation look like? Well, here it might be useful to take a look at Griffin's (1986) more holistic view of aggregation. According to Griffin (who is himself a desire theorist), "the relevant notion of aggregation cannot be simply that of summing up small utilities from local satisfactions" (p 15); "[o]ne does not most satisfy someone's desires simply by satisfying as many as possible, or as large a proportion [as possible]" (ibid., p 15). That is, he rejects the atomistic idea that the value of a person's life is a function of the values of the

---

[42]This is my reconstruction of the argument:
(1) The relational theory of pleasantness is true, i.e. an experience is unpleasant if and only if the subject has aversion A to having the experience (and so on).
(2) The aversion to suffering is (for this reason) a second-order desire: If a person has an aversion to displeasure, this implies that he has an aversion to A.
(3) On the suggestion in question (i.e. the idea that a desire should not count if one would prefer not to have it), A must be regarded as irrelevant, which in turn implies (according to Parfit) that
(4) unpleasant experiences cannot be bad for us. But this is absurd, and we should therefore reject the idea that a desire should not count if one would prefer not to have it.

The main reason why the argument is defective is that (2) does not follow from (1): If a person has an aversion to a certain displeasure, he has an aversion to having an experience to which he has an aversion, but this does not in any way imply that the object of his aversion is the first-order aversion to having the experience. For this reason, it is doubtful whether the aversion to suffering is really a second-order desire.

particular desire-fulfilments and aversion-fulfilments in it. But what is his alternative? What notion of aggregation is incorporated in his own "informed-desire account"?

Well, the basic feature of Griffin's notion of aggregation is that it is somehow based on the idea that our desires have a structure, and that this structure "already incorporates, constitutes, aggregation" (ibid., p 15). But what does he mean by "the structure of desires"?

> /.../ desires have a structure; they are not all on one level. We have local desires (say, for a drink) but also higher-order desires (say, to distance oneself from consumers' material desires) and global desires (say, to live one's life autonomously) (ibid., p 13).

In short, the suggestion is that every person's desires form a more or less coherent system (consisting of local desires, higher-order desires, and global desires), and that we must therefore conceive of our desires in a holistic way. We can then say that the more a person's preferences (regarded as a whole) are fulfilled, the better his or her life is.

But what is this supposed to mean? If we assume (plausibly) that the fulfilment of someone's preferences-as-a-whole must be in some way connected to at least some of the particular fulfilments of her particular desires and aversions, how should the idea be spelled out? Well, here it seems that what we thought was a "notion of aggregation" is really a modified desire theory that assigns some kind of special status to certain kinds of desires, viz. to global desires and higher-order desires. But what exactly does the lower status of our local first-order desires consist in? Well, they can of course be excluded altogether, or they can be counted as less relevant than our global desires, but there is also a third possibility, viz. the idea that our local first-order desires are, *on the relevant notion of strength*, simply weaker than our global desires (cf. p 206 above). This is consistent with Griffin's idea that "global desires provide, in large part, the relevant notion of the strength of desire" (ibid., p 15), and it might also explain why he claims that the structure of a person's desires "already incorporates, constitutes, aggregation". The objects of our global desires are normally much higher ranked on our preference orderings than the objects of our local desires, and (we should add) our global desires are rarely formed "on the basis of having summed local desires", or alternatively, a person's preference for one form of life over another is normally "basic", i.e. it is "not based

on /.../ other quantitative judgments" (cf. ibid., p 15). But if this is so, it seems that all we need in order to preserve Griffin's (and Parfit's) intuitions is to conceive of strength as rank in a preference ordering. This would allow for the possibility that a global desire might be outweighed by a sufficiently large number of local desires, however, and it would also imply that it is not really possible to escape the problems of the atomistic view. The question remains: If we know how many of a person's relevant desires and aversions that are fulfilled, and how strong (in the relevant sense) these desires and aversions are, how do we use this information in order to determine how well off he is (on the whole)? And if there is no plausible atomistic answer to this question (if we cannot find a plausible "atomistic" notion of aggregation), then it seems that there is no answer at all.

In short, it is doubtful whether there are any good evaluative arguments against (UD4'); it seems that all the heavy objections are of a methodological kind.

5.2. How the unrestricted desire theory can be modified in order to handle the objections above: Different kinds of modified versions of the actual desire theory

As the objections in section 5.1 have shown, the unrestricted actual desire theory is a bad version of the actual desire theory. So, how can we find a better version of the actual desire theory, a version that avoids the weaknesses of the unrestricted theory? How can the unrestricted actual desire theory be modified in order to handle the objections?

As far as I can see, the most common type of modification of the actual desire theory is restriction (or elimination), but there is also another ("weaker") way in which the theory may be modified, viz. by replacing the thesis of intensity with some claim concerning relative weights[43]. Let us take a closer look at these two types of modifications.

---

[43]It is worth mentioning that there is also a third type of modification of the actual desire theory, viz. idealization. But the result of such a modification is not an actual desire theory; to adopt such a strategy is rather to give up the actual desire theory altogether, i.e. it constitutes a rejection of (D1). We will, for this reason,

(i) *Restriction*. The unrestricted theory is modified in this way if the class of relevant desire is (somehow) restricted, i.e. if certain actual (intrinsic, now-for-now) desires and aversions are regarded as irrelevant (if they are neglected, ignored, excluded, or eliminated). That is, a restricted actual desire theory rejects (UD2), and claims (instead) that only some kinds of actual (intrinsic, etc.) desires and aversions should count as directly relevant (cf. (RD2) on pp 37-38).

It is fruitful to regard every restriction of the unrestricted actual desire theory as corresponding to some objection against the thesis of unrestrictedness. This suggests that the restrictions made by desire theorists are best classified in the same way as the objections against (UD2), i.e. into the following three groups: (a) object-oriented restrictions (restrictions to desires with the right kinds of objects, or propositional contents), (b) rationality-oriented restrictions (which excludes all desires that are, on some theory of the rationality of desire, irrational), and (c) other kinds of restrictions (e.g., restrictions which excludes all desires of which the subject is not aware, or that are not sanctioned by his higher-order desires or ideals)[44]. (A desire theory can, of course, be restricted in more than one of these ways, e.g., it can eliminate both desires with certain types of propositional contents and desires that are irrational)[45].

(ii) The second type of modification of the unrestricted actual theory that I have in mind is the type of modification we get if we reject the thesis of intensity, i.e. (UD3'), and replace it with some "non-intensity-oriented" claim concerning relative weights, i.e. a claim that gives "special status" to certain kinds of desires, but without eliminating the desires that are not given such special status. That is, a desire theory that has been modified in this ("non-intensity") way claims that is not always better for a desiring subject to have a stronger (relevant) desire

---

save this until later.

[44]It is worth pointing out that the idea that it can not be good for a person to have a desire fulfilled unless he is aware of the occurrence of the object is (strictly speaking) not a restriction claim in this sense. However, it will be discussed in this context anyway; the idea will (after all) be introduced in order to meet a certain type of objection to (UD2),

[45]Cf. Scanlon (1993), who makes a distinction between two types of restrictions, viz. (a) restrictions to desires with certain kinds of *objects*, and (b) restrictions to preferences with a certain sort of *basis*. The reason why I do not follow him here is that his talk about "basis" is somewhat unclear, and if we see what he is really after, we see that (b) is better replaced with "rationality-oriented restrictions".

fulfilled than to have a weaker desire fulfilled; that it might (in some cases) be better to have the weaker of two desires fulfilled, e.g., if (and because) it is more rational than the stronger desire, or if it is, to a higher extent, acknowledged by the subject.

Modifications of this kind can be classified in the same way as the restrictions, viz. they can be divided into (a) object-oriented modifications, (b) rationality-oriented modifications, and (c) other kinds of modifications. This is nothing but a reflection of the fact that a certain distinctions between desires (e.g., the one between rational and irrational desire) may be used in two different ways, namely (i) as a means of restriction, or (ii) as a means of determining relative weights. This is how the two cases differ: If it is claimed that only rational desires are relevant, and that all irrational desires should be regarded as irrelevant, then the distinction is used to restrict (eliminate). But if it is (instead) claimed that rational desires have more weight than irrational desires, the distinction is used to determine, in a way that is not an all-or-nothing way, how relevant a certain desire is.

This means that the different types of modifications of the unrestricted actual desire theory can be represented as follows:

	restriction	non-intensity
object-oriented	X	X
rat-oriented	X	X
other kinds	X	X

That is, there is a close correspondence between this classification of modified desire theories and the classification of objections against the unrestricted theory presented on pp 200-204[46].

My presentation of the different modifications will (as in the case of the objections against the desire theory; cf. p 204 above) be "from top

[46]The correspondence becomes even closer if we take idealized desire theories into account, i.e. if we do not restrict our attention to actual desire theories. If we do this, we get the following classification of modified desire theories:

	idealized	restricted	non-intensity
object-oriented	?	X	X
rat-oriented	X	X	X
other kinds	-	X	X

to bottom", rather than "from left to right". That is, my discussion of the different kinds of modified actual desire theories will have the following structure: First, I will focus on some possible object-oriented (or content-oriented) modifications on the unrestricted theory, then at some rationality-oriented modifications, and finally at some modifications that are neither object-oriented nor rationality-oriented (e.g., modifications which appeal to the idea that we should give more weight to desires that are sanctioned by higher-order desires, or to desires of which the subject is aware). In each of these three cases, the relevant modifications will either take the form of restrictions, where desires that are not of "the right kinds" are eliminated altogether from consideration, or they will take the form of "non-intensity" ideas of relative weight, where desires of certain kinds are regarded as *ceteris paribus* more relevant than desires of other kinds[47].

In this context, it is important to keep in mind that the main purpose of this section is to find out what possible version (modification) of the actual desire theory that is intuitively most plausible. (This theory will then be criticized later on, along with (D1)). Or more specifically, our questions are:

(i) How should the class of intrinsic now-for-now desires and aversions be restricted? According to what criteria can we determine which of a person's (intrinsic, etc.) desires that it is (nonderivatively) good for him to have fulfilled, and which of his aversions that it is bad for him to have fulfilled?, and

(ii) How do we determine which of two relevant desires that it is better for the desiring subject to have fulfilled? Should we accept (UD3') (the idea that the stronger of two relevant desires is also the more relevant one), or should we (rather) replace it with some alternative view on relative weights? If so, what view?

With these questions in mind, let us now look at some possible object-oriented modifications of the unrestricted actual desire theory.

---

[47]At this point, it is worth pointing out that there is a type of modification that has yet to be considered, viz. rationality-oriented idealizations (cf. note 46 above), where the hypothetical desires that we would have if we were rational are regarded as relevant. But we since our present concern is with actual desire theories, we will save this until later.

5.2.1. Some object-oriented modifications of the unrestricted desire theory

We already know that object-oriented modifications can take two forms, viz. (i) restrictions to desires with certain kinds of objects, and (ii) non-intensity-oriented claims concerning relative weights. All the content-oriented claims that I have come across in the literature are of the first kind, however, i.e. restriction claims, and this means that we might not have to deal with the second type of claim at all. (The reason why I have taken it into consideration is that it is a weaker kind of claim; if a certain restriction claim fails, it may be replaced with a corresponding claim concerning relative weights).

There are at least three object-oriented restriction claims which are worth investigating, viz. the following ones:

(1) The idea that only those of a person's desires and aversions which are, intuitively, about his own life should count as relevant. That is, whether those of a person's preferences that are not about his own life (that are not "personal" or "self-regarding") are fulfilled or not, this is not regarded as having any direct effect on his well-being. This is what the *Success Theory* claims.

(2) The idea that only those of a person's desires and aversions which are about his own experiences should count as relevant. This is what the *experience-oriented Success Theory* claims (this theory has already been roughly characterized in section 2.2, on pp 93-95).

(3) The claim that only global desires (that are about one's own life) should count as relevant, i.e. that all local desires should be regarded as irrelevant. This is what the *Global Success Theory* claims.

(2) and (3) are both stronger restriction claims than (1), i.e. the restrictions suggested by the experience-oriented Success Theory and the Global Success Theory are both "stronger" than the one suggested by the Success Theory. But we should not let this similarity between (2) and (3) conceal the fact that there is an enormous difference between the two theories. In fact, there are not many desires that are regarded as directly relevant by both the experience-oriented Success Theory and the Global Success Theory. The reason for this is simple: Almost all now-for-now experiential desires are desires to have (or not to have) a

certain experience, i.e. they are local desires[48].

Let us now take a closer look at these three restriction claims.

The Success Theory

On this theory, it can never be nonderivatively good for a person to have a desire fulfilled unless it is a desire about his own life. Now, in order to determine whether or not this is a plausible version of the desire theory, we must first have a more precise idea of how those of a person's desires that are about his own life should be distinguished from those that are not. Where do we ("intuitively") draw the line between the two kinds of desires?

If we assume that a person's desire that X obtains is about his own life if and only if X (the desired situation) is a part of his life, we can also formulate this question as follows: How do we determine (for a certain person) whether a certain situation (fact) is "part of his life" or not? Where do we draw the line between a person's life and "the rest", and how do we draw this line?

In most cases, it is pretty easy to determine whether a certain desire is "a desire about the subject's own life", i.e. whether a certain situation is "part of a certain person's life". As an illustration of this, here are some examples of desire-types that are clearly about one's own life: The desire to do something, or to get something done (e.g., to produce something); the desire to stand in a certain relation to a certain person, or to be with a certain person; the desires to experience certain things, or to have certain experiences; the desire to be a certain kind of person, or to live one's life in a certain way; the desire to know certain things; the desire to have certain opportunities, or to live under certain circumstances; and the desire to own certain things[49].

---

[48]A possible exception to this "rule" is (of course) the global desire to lead a pleasant life, e.g., that all one's experiences have a certain quality. In any case, the fact that most people have a strong global desire to lead a life that is happy (in the experiential sense) suggests that it is rather unlikely that the two theories will make radically different evaluations of concrete lives.

[49]This means that in this context, the term "life" is used in a rather broad sense. However, it is worth pointing out that there is also a more narrow sense of the term, and it is on this narrow usage of "life" that lives can be distinguished from (i) the persons who live these lives, (ii) the circumstances under which they are lived, and (iii) the livings (or "leadings") of them, how (in what way) they are lived. To understand the term "life" in its broadest sense, on the other hand, is to conceive of a person's life as including (i)-(iii) as well.

There are also a number of problematic cases, however, i.e. situations that are not so easily classified into the categories "part of a person's life" and "not part of a person's life". The existence of these problematic cases means that there are several possible ways in which the line between "desires about one's own life" and "desires which are not about one's own life" may be drawn. And for every such possible distinction between situations that belong to a person's life and situations that do not, we get another version of the Success Theory. Now, what we want to find is not the "true version" of theory (there is no such thing), but the most plausible one. That is, what we want to find is the line which is most plausible in the context of well-being[50]. This seems to imply that *as far as the problematic cases are concerned,* it is rather pointless to try to keep the two questions (i) "Is the situation X part of person P's life?" and (ii) "Does it have nonderivative value for P to have his desire that X occurs fulfilled?" fully separate.

There are two general types of difficulties that arise when we try to draw a sharp line between situations that are part of a person's life and situations that are not, viz. (a) problems concerning how to draw the line "in the diachronic", and (b) problems concerning how to draw the line "in the synchronic". This means that the problematic cases I have in mind can be also be divided into two groups, viz. cases that are problematic because they give rise to problems of type (a), and cases that are problematic because they give rise to problems of type (b). Let us take a closer look at these two types of problematic cases.

Richard Wollheim's (1984) notion of a life is a good example of a narrow notion of a life. This is how he distinguishes between persons, lives, and the leading (or living) of lives: "There are persons, they exist; persons lead lives, they live; and as a result, in consequence - in consequence, that is, of the way they do it - there are lives, of which those who lead them may, for instance, be proud, or feel ashamed. So there is a thing, and there is a process, and there is a product. The thing, which is a person, is extended in space, and it persists through time. Being spatial, it has spatial parts, but it does not have temporal parts. The product, which is a person's life, is extended in time, and it can be traced through space. Being temporal, it has temporal parts, but it does not have spatial parts. The process, which is the leading of a life, occurs in, though not necessarily inside, the person, and it issues in his life" (p 2).

[50]However, this does not mean that we can draw the line anywhere. There are (as we have seen) clear cases, and it is these cases which set limits to how much we are allowed to stretch the meaning of the phrase "to be part of someone's life".

About the temporal boundaries of a life

The central questions concerning the temporal boundaries of a person's life are "Should a person's own death be regarded as a part of his life?" and "Are there any post-mortem events (i.e. events which occur after a person's biological death) which should be regarded as parts of his life?" (Similar questions can of course be asked about a person's birth, about the conception, and about events which occurred before these events).

To start with the second question, I think it is rather counter-intuitive to regard any post-mortem events (including autopsies and the like) as part of a person's life. It is, after all, natural to conceive of a person's biological death as the end of his or her life, isn't it?[51] We should also remember that it has already been established (in section 4.3) that all post-mortem events can safely be ignored in a context of well-being: If all value-for-P is value-for-P-at-*t*, then things that occur after a person's death can not have any effect ("retroactive" or not) on the quality of his life. And if we also, on top of this, remind ourselves that it can never have nonderivative value for a person to have a prospective desire fulfilled (cf. pp 176-177), we can safely conclude that the most plausible version of the Success Theory regards (unlike the version put forward by Parfit (1984)) all desires for post-mortem occurrences as irrelevant: it rejects ideas like "it is good for a person to have her last wish respected", or "it is bad for a person if the post-mortem reputation that she worked so hard to establish is totally destroyed". Or alternatively put, we should (in this context) not regard desires of this type as desires about one's own life[52].

As far as our own deaths are concerned, it is hard to tell on purely semantic grounds whether they should be regarded as parts of our life or not (and whether we should think of the desire to die, or the aversion to one's own death, as preferences about one's own life). This suggests that the issue is really evaluative: the central question is how

---

[51]Here, it is important not to be misled by the fact that someone's "biographical life" - i.e. what his biography is about - often includes certain pre-natal (or "pre-conception") events (e.g., parts of the family history) as well as certain post-mortem events (e.g., what happened to his reputation, or to the cause he was working for). This does not mean that we should regard these pre-natal and post-mortem events as parts of his life, however.

[52]This is not to deny that desires of this type are often of normative relevance, however, e.g., because it is good-period that they are fulfilled.

we should conceive of death in the context of well-being, or more specifically, in the context of the desire theory. Can it, for example, ever be nonderivatively good for a person to have his desire to die fulfilled? On my own "Epicurean view" (cf. appendix C and pp 129-132), it is neither good nor bad for a person to die[53]. (This follows (but only "roughly so") from the combination of two ideas, viz. "all value-for-P is value-for-P-at-*t*" and "only things which occur at *t* can have nonderivative value-for-P-at-*t*"). But even if someone does not accept this idea as it stands, it should not be too difficult to accept the idea that the issue of death is irrelevant in the context of well-being: Who would claim that a person's death makes *him* worse off (when he is not there anymore), or that it makes his *life* worse (when there is no life there anymore that can have a value)? Moreover, the desire to die and the aversion to one's own death are always prospective, and for this reason irrelevant. For these reasons, we can conclude that it can never have value for a person to have his desire (or aversion) to die fulfilled (It may be good for a person to have his now-for-now desire to continue living fulfilled, however; cf. note 53). This means that the most plausible version of the Success Theory regards the aversion to death (or the desire to die) as irrelevant, i.e. that it does not conceive of a person's desire (or aversion) to die as a preference about his own life.

The gist of this section is really the idea that a person P's desire that X can not be a desire about his own life unless X and P are simultaneous (which does not mean that desires about one's own life can not be prospective, however). This idea does not contain anything that isn't already contained in the synchronistic idea that P's desire that X can not be relevant unless X and the desire are simultaneous, however. This brings us to the next difficulty, viz. how we should - *"in the synchronic"* - draw the line between situations that are part of a person's life and situations that are not, or alternatively, how we should distinguish those *now-for-now desires* which are about one's own life from those which are not.

---

[53]On this view, it doesn't even make sense to say that it is good (or bad) for a person to die. It makes perfectly good sense, however, to say that it is good (or bad) for a person that his life continues, just as it makes perfectly good sense to say that it is good-period (or bad-period) that he dies.

About the "atemporal" boundaries of a life

As I see it, there are two general "areas" where the "synchronic boundaries" of a life are rather difficult to determine, viz. (i) in relation to the lives of others (especially so-called "significant others"), and (ii) in relation to the "lives" of the wholes (presumably social wholes) with which a person identifies, or to which he belongs.

My life and the lives of (significant) others

I have already claimed that a person's relationships are parts of his or her life, or more specifically, that the fact that P and Q are related to each other in a certain way is a part of both P's and Q's lives. That is, the desire to be related to a certain person (or anyone) in a certain way should be regarded as a desire about one's own life, and so should (I think) the desire to do something together with someone else, or the desire to have children (or grandchildren), or the desire to occupy a certain position in other people's minds. So the question arises: Is this all the "overlap" there is between people's lives, or does the overlap in question extend beyond the relational facts just mentioned? For example, are other people's actions, experiences, achievements, sufferings, or interactions ever to be regarded as parts of my life? And can truly other-regarding desires[54] ever be regarded as being "about one's own life"? For example, is the other-regarding desire that some "Significant Other" (e.g., a child, husband, wife, lover, friend, or relative) is happy a desire "about one's own life"?

Intuitively, the answer is (with the possible exception of one's own children; cf. below) "no", and this is not changed by the fact that it sometimes makes sense to say that someone is an essential part of another person's life, nor by the fact that the suffering of a loved one may diminish one's own well-being. What happens to significant others is (typically) not part of one's own life, and the desire that something happens to someone one loves or cares about (e.g., that she will make a successful career in her field) is not a desire about one's own life.

---

[54]Where a desire is truly other-regarding if and only if its propositional content does not include any essential reference to the desiring subject. For example, P's desire that Q is doing well is not other-regarding in this sense if what P really wants is that Q is doing well as a result of his (i.e. P's) actions, and P's desire that *his* children are happy is other-regarding in this sense if "his children" can be replaced by some expression which do not contain any reference to him (i.e. to P). Cf. also note 30 on p 206.

Moreover, it would be implausible to claim that it has nonderivative value for a person to have such other-regarding desires fulfilled. Suppose I have a strong intrinsic desire that my lover will not suffer, but that she is (e.g., unknown to me) in great pain. Is this non-derivatively bad for me? I think not. The suffering of a loved one never affects one's well-being directly; if my well-being is affected by the fact that someone else is in pain, it is always indirectly, e.g., because it makes me suffer. This means that every plausible version of the Success Theory has to accept the idea that in the field of experience, there is no overlap between lives.

Does this hold in the case of one's own children as well, or should we sometimes regard desires that certain things happen (or does not happen) to one's children as desires about one's own life? For example, can it ever have nonderivative value for a parent that his intrinsic desire that his children's lives go well is fulfilled? According to Parfit (1984), the answer is "yes"; it can be bad for someone that his children's lives fail, viz. if he has a strong desire to be a successful parent, and if the children's lives fail as a result of mistakes he made as their parent (cf. p 495). I share Parfit's intuition, but I don't think it shows that it is the children's' failure *per se* that is bad for the parent. What is bad for the parent is rather the complex fact that the children's lives fail *and* this failure was caused by mistakes he made as their parent. And it is not really the desire that the children are successful that is a desire about the parent's own life, and that it is good for the parent to have fulfilled, but the desire to be a successful parent, i.e. to be a kind of person that (among other things) does not, and did not, make certain kinds of mistakes.

So, as far as I can see, there is really no reason why we should (in this context) treat children differently from other significant others, or from other people in general. No truly other-regarding desire should be regarded as a desire about one's own life, and we should never attribute prudential value to the fact that someone has such a desire fulfilled. It is true that if a person is intimately related to other people, there is a tight link between his well-being and theirs. But with the possible exception of certain relational facts, the link is psychological rather than "logical", "conceptual", or "direct". This means that the most plausible version of the Success Theory regards all other-regarding desires as irrelevant,

regardless who the other person (or creature) is[55].

My life and the "lives" of the wholes with which I identify, or to which I belong

The fact that a certain person is part of a certain whole (e.g., that he is a member of a certain club) is clearly a part of this person's life, and the desire to be part of a certain whole is clearly a desire about one's own life. This does not imply that things which happen to the whole with which I identify also happen to me, however, or that the desire that the whole to which I belong is doing well is a desire about my own life. But we may always ask whether it is. How sharp is the line between the life of a person and the "life" of a whole to which he belongs? Consider the desires that *my* team will win the series, that the company that *I* work for will make a big profit, or that the organization of which *I* am a member will grow and become more influential. It can hardly be doubted that desires like these are most often self-centered, but are they also "desires about my own life"?

In most cases of this type, the intuitive answer is "no", e.g., the fact that my company made a bigger profit this year than last year is hardly a part of my life. However, I think it hard to show on purely "semantic grounds" that facts about "significant social wholes" are *never* part of my life. It is not really necessary to settle the issue in this non-evaluative way, however. What we are after is the most plausible version of the Success Theory, and this means that when our linguistic intuitions do not take us any further, we might as well restrict our attention to the strictly evaluative side of the issue, and ask: Does it ever have non-derivative value for a person to have his (intrinsic) desires about "significant social wholes" fulfilled? Well, I think not. If a soccer player has a strong intrinsic desire that his team wins the series, and if the desire is fulfilled, the mere fact that it is fulfilled does hardly have nonderivative value for him[56]. Now, it is true that if a person identifies with a certain

---

[55]This does not mean that all intrinsic (etc.) self-regarding desires are relevant, however. Suppose that P desires that *his* friends are happy, and that the propositional content of this desire includes an essential reference to him (e.g., suppose he wants his friends to be happy because they are *his* friends, no matter who they happen to be). It is hardly nonderivatively good for P to have this desire fulfilled.

[56]Especially not if the desire is fulfilled without him knowing it, e.g., because he is in the hospital recovering from a serious injury. It is also doubtful whether he

social whole, and if he has a strong desire that this "significant social whole" is doing well, then it is likely that his own well-being is (in part) dependent on the "well-being" of the whole. But the connection between the well-being of the person and the well-being of the whole is not "logical" or "conceptual", but "causal" (in this case "psychological"): The fact that a certain significant whole is successful can only affect a person's well-being indirectly, e.g., *via* his beliefs and emotions. In short, the most plausible version of the Success Theory regards all desires about "significant social wholes" as irrelevant[57].

This concludes our formulation of the Success Theory. So, is this a plausible version of the desire theory? Well, all we can say at this point is that it is far more plausible than the unrestricted theory. The reason for this is that it is far more in line with our evaluative intuitions, and this should not surprise us: We have, after all, been making use of these very intuitions when formulating the theory! However, it remains to be seen whether the Success Theory (as it has been formulated here) is the most plausible object-oriented modification of the desire theory. To see whether it is, we have to compare it with the alternatives, the first of which is the experience-oriented Success Theory.

The experience-oriented Success Theory

The experience-oriented Success Theory is the kind of desire theory that we get if we restrict "the class of preferences to be considered" to desires and aversions about one's own experiences. On this view, all relevant preferences are preferences "for states of affairs which are /.../ presently within the experience of the person having the preference" (cf. Hare (1981), p 104). All desires that are not about one's own experiences are regarded as irrelevant, e.g., if a person has an intrinsic desire to do something, then it can not have nonderivative value for him to have this desire fulfilled. This means that the experience-oriented

---

himself would accept the idea that what is good for the team is also good for him, even if he were to remain ignorant about what befalls the team.

[57]I am aware that this might well be an expression of Modern Western Individualism, but so is the whole issue of individual well-being. The fact that there are people who "identify" with certain social wholes, and who care more about the well-being of these wholes than their own individual well-being (or the well-being of other individuals), does not mean that the individual well-being of these people is conceptually dependent on the well-being of the whole. All it means is that the issue of individual well-being is not very important to them.

Success Theory makes a much stronger restriction claim than the original Success Theory (cf. p 223 above). As Parfit (1984) puts it: "The Success Theory appeals to *all* of our preferences about our own lives. A Preference-Hedonist [experience-oriented Success Theorist] appeals only to preferences about those features of our lives that are introspectively discernible" (p 494; cf. also pp 93-95 above).

Is this a plausible theory? Or more importantly in this context, is it more plausible than the original "unrestricted" Success Theory? Well, the fact that the experience-oriented Success Theory claims that a person's well-being cannot be directly affected by things of which he is not aware may seem to speak in favour of this theory. This is (as we will see on p 266 ff.) not really the case, however: The idea that a person's welfare cannot be directly affected by things that he does not know anything about does not imply that it depends solely on his states of mind, i.e. it is also compatible with certain "state of the world" theories. There is a strong reason for regarding the original Success Theory as the more plausible theory, however, viz. the one that was given on p 128-129 (under (b)). To recapitulate: Within the framework of the desire theory, it is arbitrary to focus exclusively on a person's experiences. If the subject is regarded as sovereign in the field of his own experience, why not regard him as sovereign in his life as a whole? There is really no reason why we should attribute prudential value to the fulfilment of experiential desires only[58].

The Global Success Theory

This theory "appeals only to *global* rather than *local* desires and preferences" (Parfit (1984), p 497). On this view, only global desires about one's own life should count as relevant, i.e. if a desire is not "about some part of one's life considered as a whole, or is about one's whole life" (ibid., p 497), it should be regarded as irrelevant. Well-being is not a matter of local desire-fulfilment; how well off a person is at a certain time is (rather) a function of how satisfied he is with the life he has at

[58]At this point, it is worth noting that an unrestricted desire theorist can criticize the Success Theory in the same fashion, viz. by asking the Success Theorist what is so special about desires about one's own life that we should restrict our attention to these desires only. However, the Success Theorist can easily defend himself against this criticism by appealing to different kinds of intuitions (cf. pp 196-197 above).

that time (considered as a whole), or with its major parts[59].

The difference between global and local desires (about a person's own life-at-a-certain-time) is obviously a difference in "the relative size" of their respective objects, i.e. the difference in question is really a matter of degree. So where do we draw the line between the two kinds of desires? How "large" must the object of a desire be if this desire is to be counted as global? Well, I guess all we can do is to try to draw the line by means of examples. The desire to live a certain kind of life (e.g., a life full of pleasure) and the desire to live one's life in a certain way (e.g., in an autonomous way) are as global as they can get. The desires to have a family, to work as a doctor, or to be a brave person, are not as global, but they are (it seems) global enough to qualify as global. Examples of desires and preferences (about one's own life) that are clearly local are the desire to drink a beer, or to get rid of the anxiety that one is currently experiencing, or the preference for going to the beach rather than sitting indoors writing. (This explication of the distinction between global and local desires may not be very satisfactory, but I hope it will be sufficient for our purposes).

So, is the global Success Theory a more plausible version of the desire theory than the original "unrestricted" Success Theory? I think not, and there are several reasons for this.

Before we look at the major reasons for regarding the original Success Theory as superior to the global Success Theory, let us first note that there is really nothing that speaks for the latter theory. For example, Parfit's reason for introducing the global version of the Success Theory, and for preferring it to the unrestricted version, is (as we have already seen on pp 216-217 above) a bad reason (and this seems to apply in Griffin's case too). First, we do not really have to reject (UD4'), i.e. "the thesis of Additivity" (the "addiction objection" presented on p 216 does not really hit the idea in question), and second, even if the idea of additivity were false, this would not give us any reason to accept a global theory: It is simply wrong to view global

---

[59]Sumner (1996) sometimes seems to take it for granted that this is a feature shared by all subjectivist theories of welfare, at least all plausible theories of this type. In his terminology, a subjectivist theory is a theory which makes a person's level of well-being depend (at least in part) on his attitudes towards things like "the conditions (or circumstances) of his life", "his life as a whole", "his lifestyle", or "the significant parts (or ingredients) of his life".

theories as alternatives to "summative theories" (i.e. theories that accept (UD4')), and to believe that a global theory can somehow escape the problems that all "summative theories" have to face; at least if we assume that there are several global desires (cf. pp 217-219). This strongly suggests that there is really no need to introduce any global theory in the first place.

Moreover, the theory is intuitively plausible only as long as we regard it as a theory about how to determine how well off someone is on the whole, viz. as a theory which claims that the more someone wants his life to be the way it is, the better off he is. (Notice that if the theory is regarded in this simplistic way, it can really be regarded as an alternative to (UD4')). However, if we look at what the theory has to say about what is good and bad for a person, it entirely loses its appeal. Why is this? Well, it is simply implausible to regard all local desires as irrelevant, and the reason for this is not just that many of our local desires are connected to our global desires in such a way as to make their fulfilment important. Even if a certain local desire is not connected to some global desire in this way, it may still be good for the desiring subject to have it fulfilled. Suppose I have an intrinsic desire to drink a beer, or to get a massage, or to get rid of the pain I am in. Isn't it both absurd and arbitrary to claim that it does not have value for me to have these desires fulfilled just because they happen to be *local*? The idea that all local desires should be eliminated becomes even more absurd if we consider the fact that the difference between global and local desires is a matter of degree: Why should we eliminate a local desire that is "almost global", but not a global desire that is "almost local"?

It is not just the restriction claim made by the global Success Theory that is implausible, but also the corresponding claim concerning relative weights, i.e. the claim that global desires are, because of their "globality", more relevant than local desires. It is true that our global desires are normally more relevant than our local desires, but the reason for this is not that they are global, but that they are (in the relevant sense) stronger (cf. pp 206 and 218-219). For example, global situations such as "being happily married" or "being successful in one's work" are most often ranked much higher on people's preference orderings than more local situations such as "getting in time for the bus" or "having time to take a shower before the appointment", even though such local desires may, at times, be more intensely felt. In short,

if we have the relevant notion of strength in mind, it is unlikely that our local desires are stronger than our global ones, but even if they are (at times) stronger, this is no reason to reject the thesis of intensity[60].

This means that the global Success Theory is simply not plausible, not even if it regarded as a theory about how to determine how well off someone is on the whole, and this is not changed by the fact that it is a relatively useful theory[61]. Consider this person who tends to get caught up in details, whose many local desires are almost as strong (in the relevant sense) as his global desires. Now, assume that this person is locally frustrated but globally satisfied, that he is dissatisfied with all the details and minor events in his life, but satisfied with his life as a whole. In my opinion, it would be utterly implausible to claim that he is as well off as he can be. To conclude, we should regard the global Success Theory as far less plausible than the "unrestricted" Success Theory.

We have now shown (assuming that we have not overlooked any plausible alternatives, that is) that the Success Theory is more plausible than all its object-oriented competitors, i.e. we can conclude that the restriction claim it makes is the most plausible object-oriented restriction claim there is. The theory is hardly the most plausible version of the desire theory, however, and the reason for this is that it is vulnerable to the rationality-oriented objections on p 207 ff., and perhaps also to the objections on pp 211-214. These objections must be taken into account, however, and this suggests that in order to arrive at the most plausible version of the desire theory, we must modify our theory further, viz. (among other things) in a rationality-oriented way. But we will keep the idea that only desires that are about one's own life can be relevant, i.e. we will assume that the best version of the desire theory is

---

[60]This suggests that we need not invoke the global Success Theory in order to explain why we tend to view global satisfaction with one's own life as a necessary condition for having a good life. If we combine the fact that our strongest desires about our own lives are often global desires with the thesis of intensity, we get an equally good explanation.

[61]The reason for this is twofold: First, the fact that the theory does not take so many desires into account makes it easier to use in practice. (One way in which this can be done is this: A number of life areas (or domains) are determined, and the subject can then assess his own level of satisfaction in these domains). Second, the fact that our global desires are almost always much stronger (in the relevant sense) than our local desires makes it likely that the comparisons of well-being that are based on the global Success Theory will often coincide with the comparisons that are based on the original Success Theory.

a modification of the Success Theory. So let us now look at how this theory can be further modified.

5.2.2. Rationality-oriented modifications of the Success Theory[62]

So, I will assume that at least some of the rationality-oriented objections in section 5.1.2. are valid, and that the most plausible version of the desire theory is (for this reason) a rationality-oriented modification of the Success Theory. There are many possible rationality-oriented ways in which the Success Theory may be modified, however, and it is far from obvious which of these possible modifications we should prefer. To clarify the problem of which modification (or modifications) that should be selected, we can divide it into two interrelated questions:

(1) If there is at least one valid rationality-oriented objection to the unrestricted desire theory, then there is at least some rationality-oriented claim concerning relative weights which has to be accepted as valid. That is, we have to accept that those of a person's desires which are (in the relevant sense) rational are more relevant than those which are irrational, or alternatively put, that it is *ceteris paribus* better for a person to have a rational desire fulfilled than to have an irrational desire fulfilled. So the question arises: Should we also accept the corresponding restriction claim, i.e. should we conceive of all desires that are irrational (in the same relevant sense) as totally irrelevant?[63]

(2) Every rationality-oriented desire theory is (obviously) based on

---

[62]It is important to keep in mind that in what follows, we will take the Success Theory for granted. We should also remind ourselves that the aim of the investigation is still to find the most plausible version of the desire theory, i.e. we will not yet put (D1) into question.

[63]Here, we might remind ourselves that there is also a third kind of rationality-oriented modification, viz. the idealization (where the hypothetical desires that we would have if we were rational are also regarded as relevant). All idealizations are rejections of (D1), however, and they do therefore not belong in this context (cf. also note 43). It is also worth mentioning that all the rationality-oriented modifications I have come across in the literature are either restrictions or idealizations (which are both "all-or-nothing claims"): I have never come across any rationality-oriented claims concerning relative weights. My reason for including claims of this kind is that they are weaker than the corresponding restriction claims; e.g., if a certain restriction claim fails, it may be replaced with a corresponding claim concerning relative weights.

some idea of what it is for a desire to be rational (or irrational), i.e. of what the rationality (or irrationality) of a desire consists in[64]. Now, there are (as far as I can see) at least five different conceptions of rational desire, conceptions which give rise to five different "senses" in which a desire might be rational or irrational. These conceptions can be (roughly) characterized as follows:

(i) On *Hume's theory*, a derived desire is rational if and only if it is based on true beliefs (our underived desires are, on this view, neither rational nor irrational).

(ii) *The informed theory* claims that a desire is rational if and only if it is informed, where a desire is informed if and only if it is "formed by appreciation of the nature of its object" (cf. Griffin (1986), p 14)[65].

(iii) *The deliberative theory* claims that a person's actual desire is rational if and only if it would survive a process of ideal deliberation, e.g., if he would still desire it "if he knew the relevant facts, was thinking clearly, and was free from distorting influences" (cf. Parfit (1984), p 118).

(iv) On *the genetic theory*, a desire is rational if and only if it has the right kind of causal history, e.g., if it has not been shaped by

---

[64]That is, the focus is on the rationality of particular desires and aversions (considered more or less in isolation), and not how we should determine whether a certain *set* (or system) of preferences is rational or irrational (where it is assumed that such a set can be irrational, even if the desires in the set are not in themselves irrational, e.g., because they are inconsistent with each other). However, this does not mean that a desire theorist can ignore the question of how to determine whether a certain set of preferences is rational or irrational; first, he could use an answer to this question to modify (UD4') in a more holistic direction, and second, he might (for some reason) try to capture the rationality of particular desires in terms of the rationality of sets of desires (e.g., because he believes that a particular desire can not be fully rational unless it is an element in a rational set).

[65]It is worth noting that this theory is my own construction, i.e. as far as I know, there is no one who has claimed that a desire is *rational* in virtue of being informed. So, what are my reasons for constructing such a theory? Well, first, I think it is fruitful to regard Griffin's informed desire theory as a type of rationality-oriented desire theory, and second, the informed theory (regarded as a conception of rational desire) is a kind of modern Humean theory, and it can be seen as a kind of natural step on the way to the deliberative theory. (It may also be added that I used to believe that the informed theory contained nothing that wasn't already contained in the other four conceptions, but that I have changed my mind on this point).

"irrelevant causal processes" (e.g., by manipulation).

(v) On *the intrinsic theory*, there are desires that are intrinsically rational and irrational, where an intrinsically rational desire is a desire that is rational because of its content, e.g., in virtue of the fact that its object is (in some objective sense) worth desiring[66].

Now, the fact that there are several different conceptions of rational desire means that we must ask ourselves which of these conceptions that is of most interest in this particular context. Which conception of rational desire gives, if incorporated into a rationality-oriented desire theory, rise to the most plausible theory[67]?

The problem can also be formulated in a somewhat different (and perhaps simpler) way, viz. as follows: Every rationality-oriented modification of the Success Theory either takes the form of a restriction claim or a claim concerning relative weights, and there are (moreover) five conceptions of rational desire on which such claims can be "based". This means that there are ten possible modifications of this type. These possibilities can be schematically represented as follows:

	restriction claim	claim conc. rel. weights
Hume's theory	x	x
the informed theory	x	x
the deliberative theory	x	x
the genetic theory	x	x
the intrinsic theory	x	x

Now, our question is simply which of these ten possible modifications that are (I assume that it might be more than one) plausible.

The investigation will (as usual) proceed "from top to bottom" rather than "from left to right", i.e. we will consider one conception of rational desire at the time, and in connection with each conception, we will try

---

[66]There are (of course) alternative (and perhaps also better) classifications, but I think the present classification is good enough for our purposes.

[67]That is, the question is not which of the conceptions that is true, but what conception of rational desire that makes a rationality-oriented desire theory most plausible. However, the fact that our focus is on the issue of evaluative relevance (or plausibility) does not mean that we have no interest in things like which (if any) conception that is most "descriptively relevant" (e.g., most in line with ordinary speech).

to find out whether any of the two modifications which are based on this conception is plausible. That is, for each conception of rational desire, two questions will be asked, viz. (i) Is the restriction claim that is based on this conception plausible, i.e. should we regard all desires that are (on this conception) irrational as irrelevant?, and (ii) If not, should we (at least) accept the corresponding claim concerning relative weights, i.e. should we conceive of desires that are (on this conception) rational as more relevant than desires that are not? If a certain desire D1 is rational (in this sense) while another desire D2 is not, can it (then) be better for the desiring subject to have D1 fulfilled, even if D2 is stronger?[68]

Let us start with Hume's theory.

Rationality-oriented modifications based on Hume's theory

So, is it reasonable to regard all desires that are irrational in the Humean sense as irrelevant? If not, should we regard those desires which are (on Hume's theory) rational as more relevant than those which are not?

To get in a better position to answer these questions, let us first take a closer look at Hume's (1739-40) conception of rational desire. This conception consists of two central claims, viz. (a) the idea that only derived desires that can be classified as rational or irrational, and (b) the idea that a derived desire is irrational if and only if it is based on one or several false beliefs, and rational if and only if it is based on true beliefs[69]. This is Hume's own famous formulation of the theory:

> Where a passion [desire, preference] is neither founded on false suppositions, nor chuses means insufficient for the end, the understanding can neither justify nor condemn it. 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledge'd lesser good to my greater, and have a more

[68]I will, in each case, start with the restriction claim. My reason for this is twofold; it is stronger, and it is the type of claim we find in the literature.

[69]In this context, Hume does not distinguish between true and rational (e.g., justified) beliefs, or between false and irrational beliefs.

> ardent affection for the former than the latter. /.../ In short, a passion must be accompany'd with some false judgement, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment[70] (Book II, part III, section III, p 463).

The "false suppositions" (or "judgements") which Hume himself had in mind are of two different kinds, viz. "supposition[s] of the existence of objects, which really do not exist" and false "judgements of causes and effects" (ibid., p 463). This suggests that there is, on Hume's view, only one way in which an *intrinsic* desire may be irrational, viz. if it is "founded on the supposition of the existence of objects, which really do not exist". But here, we can (in the Humean spirit) add that this is not the only type of false belief that can make an intrinsic desire irrational, e.g., an intrinsic desire may also be irrational if it is based on a false belief of the subsumption type, or perhaps of the part-whole type (cf. p 184 above), or if it is based on a mistaken attribution of properties to a certain thing[71].

But how should phrases like "based on", "founded on", and "accompany'd with" be understood here? If a desire D is based on a belief B, what kind of relation is it that holds between D and B? For example,

---

[70]That is, it is (on the Humean view) really only beliefs which can be criticized on rational grounds. However, this does not mean that Hume's conception of rational desire is a "universalist conception": The fact that "if my belief that X is true, so is yours" does not have any implications for the rationality of our desires. In fact, Hume's theory allows for the possibility that my desire that X is rational while yours is not.

[71]It can hardly be doubted that our intrinsic desires-that can be based on false beliefs (e.g., the intrinsic desire to own a certain thing can be based on false beliefs about the intrinsic properties of the thing). However, it is worth asking whether our intrinsic desires-that can ever be based on false beliefs *about their objects* (assuming that the objects of our intrinsic desires-that are really situations-under-"intrinsic"-descriptions (cf. pp 165-166 above) rather than situations *per se*, and that we can (for this reason) not really distinguish a situation (*qua* object of desire, that is) from its descriptions). Well, I think they can, at least as long as we do not incorporate the belief on which an intrinsic desire is based into the propositional content of the desire. Suppose that P intrinsically desires to go jogging because he believes that it is relaxing. On my view, the object of this desire is the fact that P is jogging (period), and not that he engages in this jogging-*qua*-relaxing-activity. And it surely makes sense to say that the desire to jogging (period) is based on the false belief that jogging is relaxing (but it does not necessarily make sense to say that the desire to go jogging-*qua*-relaxing-activity is, in this case, based on a false belief).

should "based on" be understood as "derivable from" or "(in fact) derived from", i.e. is the relation between D and B purely logical, or is it only partly logical (and partly causal)?

Now, it is hardly plausible to claim that a certain desire of P's is irrational in virtue of being *derivable* from some false belief that P has. So, is it more plausible to claim that a desire is irrational if and only if it is *actually derived* from some false belief? Well, there is a problem with this view too: What if I have this desire which is actually derived from some false belief, but which is derivable from my true beliefs (it is, after all, possible to arrive at the right answer in the wrong way); should we really regard this desire as irrational? I am not quite sure about this, but I think not. On my view, the most plausible Humean view is a kind of compromise between the two views, and it can be formulated as follows: A person P's desire is irrational if and only if (i) it is actually derived from false beliefs, *and* (ii) it is not derivable from P's (actual) true beliefs. That is, if we regard all desires that are not irrational as rational, this would imply that those of a person's desires which are derivable from his (actual) true beliefs are all rational.

To return to the question of relevance: Should we regard all (intrinsic) desires that are irrational in this sense as irrelevant, or does it sometimes have nonderivative value for a person to have such a desire fulfilled? Suppose that John has an intrinsic Hume-irrational desire to spend time with Paul. Suppose also, to put all hedonistic and other instrumental considerations aside, that it does not make him feel good (or that it is useful for him) to have the desire fulfilled. Can it, in this case, be good for him to have the desire fulfilled? Well, I think so. Consider the following case: John's desire to spend time with Paul is actually derived from false beliefs (e.g., the belief that Paul is a funny bloke), and it is not derivable from the true beliefs that John actually has. However, Paul is an interesting person, and he thinks that John is a great guy, and if John would come to believe this, he would still want to spend time with Paul, regardless of whether Paul is funny or not.

This suggests that we should reject the rationality-oriented restriction claim which is based on Hume's conception. What matters is not whether our desires are derivable from the true beliefs we actually hold, but whether they are (so to speak) "derivable from the truth", i.e. whether we would have them if we were informed about the relevant

facts[72].

Let us now turn to Griffin's informed desire theory, which is (if viewed as a rationality-oriented modification) based on a modern "version" of Hume's theory, viz. the informed theory.

Griffin's informed desire theory

On this theory, "'utility' is the fulfilment of informed desires, the stronger the desires, the greater the utility" (Griffin (1986), p 14). That is, only informed desires are counted as relevant[73]. Is this a plausible view: can it never have nonderivative value for a person to have an uninformed desire fulfilled?

So, what is it (on Griffin's view) for a desire to be informed (or un-

---

[72]Where the assumption is that our underived desires would not change if we would become more informed. That is, the whole focus is still on derivability (which is a logical relation): the fact that some of our underived desires are causally dependent on what we believe (and that confrontation with facts and logic can also, for this reason, have a causal effect on what we desire) has not yet been taken into account.

[73]This means that Griffin's informed desire theory is (at least in part) a restricted theory. Is it also an idealized (or ideal) theory, i.e. does it also regard as relevant the desires people would have if they were rational or informed? Scanlon (1993) seems to think so, when he attributes to Griffin the idea that "the quality of people's lives depend only on the fulfilment of those desires that *they would have* if they 'appreciated the true nature' of the objects of those desires" (p 187, my italics). Now, this interpretation is clearly consistent with some of the things Griffin (1986) writes - e.g., with the claim that "what must matter for utility will have to be, not persons' actual desires, but their desires in some way improved" (p 10) - but it is, nonetheless, wrong. To see why we should conceive of Griffin's informed desire theory as a restricted theory rather than as an idealized (or ideal) theory, consider the following passage from Griffin (1986): "At this point, an obvious move is to say that desires count towards utility only if 'rational' or 'informed'. 'Utility', we might try saying, is the fulfilment of desires that persons would have if they appreciated the true nature of their objects. *But we shall have to tone this definition down a bit.* Although 'utility' cannot be equated with actual desires, *it will not do, either, simply to equate it with informed desires.* It is doubtless true that if I fully appreciated the nature of all possible objects of desire, I should change much of what I wanted. But if I do not go through that daunting improvement, yet the objects of my potentially perfected desires are given to me, I might well not be glad to have them; the education, after all, may be necessary for my getting anything out of them. This is true, for instance, of acquired tastes /.../. *Utility must, it seems, be tied at least to desires that are actual when satisfied.* (Even then we should have to stretch meanings here a bit: I might get something I find that I like but did not want before because I did not know about it, nor in a sense want now simply because I already have it; or I might, through being upset or confused, go on resisting something that, in some deep sense, I really want.)" (p 11, my italics).

informed)? Well, first, an informed desire can be characterized negatively as a desire that is not faulty, or defective, in any of the following three ways: (i) It does not rest on mistakes of fact[74], (ii) nor on logical mistakes (like "confused, irrelevant, or question-begging" reasoning). (iii) The third, "conceptual", defect is harder to grasp, and the reason for this is (I think) that it is not attributed to individual desires, but to whole sets of desires. This is suggested in the following passage:

> Sometimes desires are defective because we have not got enough, or the right, concepts. Theories need building which will supply new or better concepts, including value concepts. For instance, it is easy to concentrate on desires to possess this or that object, at the cost of more elusive, difficult-to-formulate, desires to live a certain sort of life (ibid., p 12-13).

(iv) Informed desire is also positively characterized, as a desire "formed by appreciation of the nature of its object"[75], and we are also told that informed desires are based on an understanding of "what properties things and states of affairs have, and we must put our desires through a lot of criticism and refinement to reach this understanding" (ibid., p 14).

In order to make this conception of informed desire precise enough, we need to know at least two things, viz. (a) whether "the appreciation of the nature of the object" to which Griffin refers includes correct value judgements or not, and (b) when it is claimed that informed desires do not rest on mistakes of fact (etc.), or that an informed desire is formed by appreciation of the nature of its object, what do the phrases like "rest on" and "formed by" mean here; should they be interpreted as "derivable from", "derived from", "caused by", or some combination of these? (cf. pp 240-241 above). Here, I will simply assume (a) that the "appreciation" referred to is value-free[76], and (b) that the

---

[74]I take this to mean that all desires that are irrational in the Humean sense are also uninformed. The reverse is not true, however.

[75]At this point, Griffin continues by saying that an informed desire "includes anything necessary to achieve it [the object]". Isn't this a very odd idea? And what does "the object" refer to here? Well, it is most probably not the object of the desire (i.e. some situation), but some component in the situation (i.e. some thing).

[76]This would imply that Griffin's informed desire theory is a "relativist" rather than "universalist" conception of rational desire. However, it is not just that the fact that I have an informed desire that X does *not* imply that your desire that X is informed too; the theory also seems to allow for the possibility that I have an

phrases "based on" and "formed by" should not be understood in a strictly causal sense; the idea that a desire can be irrational in virtue of being caused by (or causally dependent on) false beliefs is an essential part of the deliberative theory, and it will therefore be discussed in connection with this theory. So, should "rest on" be interpreted as "derivable from" or "derived from"? Well, it seems clear that Griffin has the latter alternative in mind, i.e. I take (i) and (ii) to mean that a desire can not be informed unless it is actually derived from actual true beliefs, and I take (iv) to require that these true beliefs are about "the object", and that there are no relevant facts about the object of which the subject is ignorant[77].

So, now that we have some idea of what it is that makes a desire informed (uninformed): Is it plausible to regard all desires that are uninformed in Griffin's sense as irrelevant? Or alternatively put (considering that there are several different ways in which a desire can be uninformed): (i) Are all desires which are derived from "mistakes of fact" (i.e. from false beliefs) irrelevant? (ii) Does the fact that a desire is causally dependent on logical mistakes (like "confused, irrelevant, or question-begging" reasoning) make it irrelevant? (iii) Does the fact that a desire is "conceptually defective" give us a reason for regarding it as irrelevant (or less relevant)? (iv) Does the fact that a desire is not formed by an "appreciation of the nature of its object" make it irrelevant?

We have already seen (in connection with the Humean theory) that the answers to (i) and (ii) are both "no". The fact that a desire is actually derived from false beliefs, or that it has actually been arrived at in a faulty manner, does not make it irrelevant; it is, after all, possible to arrive at the right answer in the wrong way. What matters is whether a desire is (properly) derivable from more fundamental desires and "the truth". (iii) is probably (in this context, that is) an unintelligible question, and it can therefore be ignored[78]. And as far as (iv) is concerned, it is

---

informed desire that X while you have an informed aversion to "the same thing".

[77]This suggests that all informed desires would survive a confrontation with all facts about their respective "objects". However, this does not mean that every informed desire would survive a confrontation with all facts: we cannot conclude that if a desire is informed, it is also deliberatively rational.

[78]To be more precise: Griffin's conceptual idea is (because of its holistic character) irrelevant in relation to the questions like "what desires is it good for a person to have fulfilled?" (but only on the assumption that the rationality of a particular

far too strong. The fact that a desire is derivable from more fundamental desires and "the whole truth about its object" is sufficient to make it relevant.

To conclude, it seems that we still stand on the same point we stood on after our discussion of Hume's theory. That is, the only rationality-oriented restriction we have arrived at so far is the idea that a derived desire is irrelevant if it is not derivable from more fundamental desires and "the whole truth about its object" (which implies that the desire could not have existed were it not for false beliefs, faulty reasoning, or ignorance[79]).

Rationality-oriented modifications based on the deliberative theory

Let us now ask ourselves whether it is plausible to regard all "deliberatively irrational" desires as irrelevant, and if it is not, whether we should (at least) accept the weaker claim that desires that are rational in the deliberative sense are more relevant than desires that are not.

To be able to answer these questions, we need a better understanding of the deliberative theory of rational desire. On this theory, a person's desire that X is deliberatively rational if and only if he *would* desire that X after having had undergone a process of ideal deliberation, and a person's desire is deliberatively irrational if and only if he would not have it under these ideal circumstances. If we restrict ourselves to a person's actual desires, we get: A person's actual desire is deliberatively rational if and only if it would survive a process of ideal deliberation, and it is deliberatively irrational if and only if it would not

---

desire is independent of the rationality of the set of which it is an element). However, Griffin's idea may well be of relevance in relation to the more holistic "how do we determine how well of a certain person is?". (But how? If the idea is simply that some people do not want all the things they "should want", it can be expressed in a much simpler way, e.g., by referring to things that are (in some objective sense) "worth desiring"). It is also worth noting that the idea in question is much more in line with idealized desire theories than with actual desire theories (it is not clear whether it can be used to criticize (D1), however).

[79]This is the most interesting difference between Griffin's theory and Hume's: Where Hume is focusing solely on false beliefs, Griffin gives just as much weight to ignorance. But isn't ignorance just a special case of false belief? No, it isn't (it seems possible to not have any beliefs at all about certain things), but the reverse is of course true.

survive such a process[80].

So, what is this "process of ideal deliberation" to which the deliberative theory refers? Well, if we look at how this "process" has been characterized by different writers, we will see that it can (roughly) be viewed as a kind of mixture between cognitive activities and cognitive states: the idea seems to be that what it is rational for a person to desire is what he would desire if he would be in a certain cognitive state *and* if he would engage in a certain kind of cognitive activity. The relevant cognitive state (or states) has been described in several different ways, e.g., as "knowing the facts" (Parfit), "having all the relevant factual information" (Harsanyi), and "being correctly or fully informed about one's circumstances" (Gauthier)[81]. And the relevant cognitive activity has been described as e.g., "reasoning with the greatest possible care" (Harsanyi), "thinking clearly, [in a way that is] free from distorting influences" (Parfit), and "full and careful reflection" (Gauthier)[82].

In order to get an even better grasp of the deliberative theory, it may be fruitful to take a closer look at the perhaps most well-reasoned version of this theory, viz. the one developed by Brandt (1979). He writes:

> I shall call a person's desire, aversion, or pleasure 'rational' if it would survive or be produced by careful 'cognitive psychotherapy' for that person. I shall call a desire 'irrational' if it cannot survive compatibly with clear and repeated judgements about established facts. What this means is that rational desire (etc.) can confront, or will even be produced by, awareness of the truth; irrational desire cannot (p 113).

---

[80]That is, a desire is (on this view) always rational or irrational relative to some individual or other. This means that if we assume (plausibly) that people would remain different even after having had undergone a process of ideal deliberation, we can conclude that the deliberative theory allows for the possibility that it is rational for one person to desire a certain thing, but irrational for another person to desire the same thing. In short, the theory must be conceived of as a "relativist" conception of rational desire.

[81]Here, we may also add Griffin's "appreciating the true nature of the object" and "realizing fully what is at stake". However, we can (in this context) ignore Nussbaum's (1990) talk about what someone would desire if her "education and knowledge of alternatives were above the threshold of what is required for practical reason and choice" (note 32, p 245); there is, after all, nothing *ideal* about this cognitive state.

[82]This might not be a complete characterization of the "ideal process", however, i.e. it might be necessary to add a third element, e.g., that the relevant information is presented in an "ideally vivid way" (cf. below).

So what is this *cognitive psychotherapy* to which Brandt refers? Well, it can be characterized as a

> process of confronting desires with relevant information, by repeatedly representing it, in an ideally vivid way, and at an appropriate time. /.../ [T]he process relies simply upon reflection on available information /.../. It is *value-free reflection* (ibid., p 113).

That is, the idea is that a person's desire is rational if and only if he would have it if his "total motivational machinery were fully suffused by available information", i.e.

> if relevant available information registered fully, that is, if /.../ [he] repeatedly presented to /.../ [himself], in an ideally vivid way, and at an appropriate time, the available information which is relevant in the sense that it would make a difference to desires and aversions if /.../ [he] thought of it (ibid., p 111).

This means that the deliberative theory can be regarded as a special version of the general view that a desire is rational if it has been influenced by facts and logic as much as possible (cf. ibid., p 113). That is, the theory is (just like the other two theories above) based on the idea that our desires can be influenced by our beliefs. But where these theories restrict their attention to one kind of influence only, viz. to a type of "derivation" which involves logical dependence, the deliberative theory also attaches great weight to the idea that our desires may also be "purely" causally dependent on our beliefs. And the beliefs on which a certain desire is causally dependent need not be beliefs about its object, it may also be beliefs about other things (e.g., the beliefs the subject has about himself). In fact, the deliberative theory is (in a way) based on this "insight", or more specifically, the insight that the desires which are causally dependent on careful (value-free) reflection on relevant information are (as a rule) *superior* to the desires which are causally dependent on false beliefs and/or confused reasoning. But it is important to note that the theory does not claim that *every* single desire which is, *in fact*, causally dependent on false beliefs and/or confused reasoning is therefore irrational; what matters is (as we have seen) whether it *would survive* a confrontation with clear and repeated judgements

about established facts[83].

The perhaps most problematic feature of the deliberative theory is that it does not offer a satisfactory account of what it is that *makes* a certain desire rational (or irrational). For example, it may well be true that a desire is irrational if and only if it would not survive a process of ideal deliberation, but it is hardly the case that a desire is irrational *in virtue of* the mere fact that it would not survive such a process. So, what is it that makes a deliberatively irrational desire irrational? Well, it seems clear that if a desire would extinguish in cognitive psychotherapy, it would do so because of some other property that it has, e.g., because it is dependent on false beliefs (or ignorance about certain things). So why don't we simply regard all properties of this kind as "irrationality-making properties"? On this view, there are a number of "irrationality-making" features of desires such that every deliberatively irrational desire is irrational in virtue of possessing some such feature, and such that if a certain desire has such a feature, it is likely (but not certain) that it would (for this particular reason) not survive a thorough confrontation with facts and logic. But what properties of desires can reasonably be regarded as "irrationality-making" in this sense?

Even though Brandt does not give any explicit answer to this question, he gives several examples of (kinds of) desires and aversions that would most probably *not* survive cognitive psychotherapy, and this may give us some fruitful ideas of *why* certain desires would extinguish in such a process.

What some of the desires and aversions that are (on Brandt's view) "unfit" to survive cognitive psychotherapy have in common is (roughly) that they have originated in the "wrong ways"[84]. Examples of such desires are (1) desires and aversions that have developed from generalization from untypical examples, i.e. "from familiarity with samples of liked/disliked situations but from *untypical* samples, or in an untypical

---

[83]And since it would be absurd to claim that only desires which are derived (or derivable) from true beliefs would survive such a confrontation, we can see why the deliberative theory has to reject the Humean idea that only derived desires can be classified as rational or irrational. However, it may well be the case that all (or most) intrinsic desires that are irrational in the Humean sense are also deliberatively irrational.

[84]This shows that there is an intimate connection between the deliberative theory and the genetic theory. This connection will be further discussed on pp 255-256 below.

context" (ibid., p 120); and (2) exaggerated desires and aversions produced by "an early and prolonged deprivation of something wanted - enough for discomfort and anxiety to be involved", e.g., abnormally high and perhaps insatiable desires for attention, commendation, admiration, or company (cf. ibid., p 123)[85].

It is probably true that desires which have been shaped in any of these ways would, in many cases, not survive cognitive psychotherapy, but this is merely an empirical fact, and it does not allow us to conclude that any of these desires would be extinguished *in virtue of* having been shaped "in the wrong way". Moreover, desires with the wrong kind of causal history could well survive cognitive psychotherapy, and the idea that a desire may be "deliberatively irrational" in virtue of its causal history should therefore be rejected. Instead, we should accept the more plausible view that desires which are shaped in certain ways will often get (as a result of this) certain other defects, and it is in virtue of these defects that they would not survive a confrontation with facts and logic.

So, if we disregard the desires which have been shaped "in the wrong way", what other examples of kinds of desires which are "unfit" to survive cognitive psychotherapy can we find in Brandt (1979)?

(1) Well, first, there are the desires (and aversions) which Brandt calls *artificial*. These are the desires and aversions that "could not have been produced naturally", that "could not have been brought about by experience with actual situations which the desires are for and the aversions against" (p 117)[86]. As far as I can see, there is no reason why we should not regard the property of "artificiality" as an "irrationality-making" feature[87].

(2) There is (of course) also the idea that desires can be unfit to survive in virtue of being causally dependent on false beliefs (or "ignorances"). Now, if the phrase "causally dependent" is interpreted as

---

[85]But is it really so unlikely that desires of this kind would survive cognitive psychotherapy? Well, this is what Brandt believes. "[I]f a person brought to consciousness the connection of early deprivation with his intense desire (=insight), the abnormally strong desire would [on his view] abate" (ibid., p 124).

[86]Brandt also adds that most of these desires are (in fact) produced by "observation of the attitudes and values of other persons - parents, teachers, or peers, not to mention films and television" (ibid., p 116). However, it is important to note that this is not what makes them artificial.

[87]The idea that desires that are artificial (in this sense) would extinguish is not very informative, however; it may even be tautological.

"maintained by" rather than as "produced by", it is not just plausible to assume that those of a person's desires that are causally dependent on false beliefs would extinguish in cognitive psychotherapy, but also that they would do so *because* the person has continued to have them because of certain false beliefs (or "ignorances"), e.g., about the object of the desire, or about himself. That is, it seems that the property of being causally dependent on (maintained by) false beliefs is a truly "irrationality-making" feature[88].

So, are there any possible irrationality-making features that we have overlooked? Well, the only property I can think of is the property of being intrinsically irrational in a certain way, viz. of being directed towards an object that is (in some objective sense) "in no respect worth desiring", or even "worth avoiding". A deliberative theorist would deny that there are such desires, however, and they can therefore be ignored in this context. Suffice it to say that it is not unlikely that most of the desires that an intrinsic theorist classifies as intrinsically irrational would extinguish in cognitive psychotherapy, but this is not to say that they would do so because of their alleged intrinsic irrationality (especially not if there is no such thing).

This concludes our exposition of the deliberative theory. Let us now return to the issue of relevance, and ask whether we should accept any of the two possible modifications (of the Success Theory) which are based on this conception. The first question was: Should we regard all (intrinsic) desires that would not survive a process of ideal deliberation as irrelevant?

Suppose that John has a strong, intrinsic desire to write a book. Suppose also that this desire would not survive a confrontation with facts and logic. Does this allow us to conclude that it would not have nonderivative value for John, as the person he is now, to write the

---

[88]So, what does Brandt have to say on what kinds of (intrinsic) desires that would be likely to *survive* or be *produced by* cognitive psychotherapy? Not much: The only kind of desires and aversions that he is reasonably sure would be accepted as rational are desires for what is natively liked (e.g., dozing when tired, eliminating, eating good food, having novel stimuli, and tasting something sweet), and aversions for what is natively disliked (e.g., being in pain, being too cold or too hot, being unable to breathe, having to stay awake when sleepy, being bored, stubbing one's toe, and receiving an electric shock); cf. ibid., pp 130-131. This does mean that only "baby-wants" can be rational, however, e.g., Brandt would most probably accept the view that those of our wants which are autonomous, or natural, or informed, are also "fit to survive".

book and have the desire fulfilled? I think not, but neither can we conclude that it would be good for John to write the book. We simply don't have enough information to answer the question. In order to determine whether John's desire is relevant or not, it is not enough to know *that* it would not survive cognitive psychotherapy, we also need to know *why* it would not survive. The more general point I want to make can be expressed as follows: The mere fact that a desire would not survive cognitive psychotherapy does not *make it* irrelevant (or less relevant). If a certain deliberatively irrational desire happens to be irrelevant, it is not in virtue of its deliberative irrationality, but for some other reason, i.e. because of some other property that it has[89]. That is, the connection between irrelevance and deliberative irrationality is contingent rather than necessary.

So, if a deliberatively irrational desire happens to be irrelevant, in virtue of what is it irrelevant? Well, I think it can be assumed that the reason why a certain irrelevant deliberatively irrational desire is irrelevant tends to coincide with the reason why it is deliberatively irrational (why it would not survive cognitive psychotherapy). Or alternatively put, I believe that *if* a certain deliberatively irrational desire happens to be irrelevant, then the property which makes it irrelevant is most probably the same property which makes it "unfit" to survive cognitive psychotherapy. This does not imply that all "irrationality-making" properties are also all "irrelevance-making" properties, however, but only that the properties which make deliberatively irrational desires irrelevant are to be found among the properties which make these desires deliberatively irrational in the first place. In fact, it seems plausible to assume that there are "irrationality-making" properties which are not (at the same time) "irrelevance-making". A possible example of such a feature is the property of being causally dependent on false beliefs (or ignorance) about one's own history. A desire which possess this property would probably not survive a confrontation with (psychological) facts about how it originated, but it is doubtful whether we should regard it as irrelevant.

So, which of the irrationality-making properties mentioned above can (and should) also be regarded as "irrelevance-making"? Well, the only

---

[89]This idea must be carefully distinguished from the (similar) idea that every deliberatively irrational desire is *irrational* in virtue of some "irrationality-making" property that it has.

serious candidates I can think of are (1) the property of artificiality, and (2) the property of being causally dependent on false beliefs (or ignorances).

(1) So, should we regard all artificial desires as irrelevant? Does the fact that a certain desire "could not have been produced naturally" make it irrelevant? Well, it seems clear that most (or all) *derived* artificial desires have to be regarded as irrelevant: This seems to follow from the restriction claim which has already been adopted, viz. the idea that a desire is irrelevant if it is not (in a strict sense) derivable from more fundamental desires and the whole truth about "its object" (i.e. the actual situation which the desire is for). But what if we have *underived* desires in mind? Well, in this case, the artificiality of a desire seems to consist in its causal dependence on false beliefs or ignorance, and the idea that such desires should be regarded as irrelevant can therefore be discussed under (2).

(2) So, should we accept the idea that if a desire is causally dependent on false beliefs (or ignorance), then it can never have nonderivative value for the desiring subject to have the desire fulfilled? Let us return to John's desire to write a book, and let us assume that the reason why this desire would not survive cognitive psychotherapy is that it is causally dependent on false beliefs. Does this mean that it would not be good for John to have the desire fulfilled? Well, this may well depend on what the false beliefs are about. If John's desire is dependent on false beliefs about the activity of writing, or about his own ability to write, it is (I think) irrelevant, but what if it is dependent on his ignorance about how the desire came into existence in the first place? Would this make it not-good for John to have the desire fulfilled? I think not. My tentative view can be formulated as follows: The fact that a person's desire is causally dependent on some false belief (or state of ignorance) can make it irrelevant, but only if (i) the desire is underived (if it is derived, it is the restriction claim above which holds), *and* (ii) the false belief (or ignorance) is (intuitively) "about the object", e.g., if the person desires that X because he is ignorant about Z, then Z is (somehow) a part of the situation X.

If we, on top of this, conceive of *all* underived desires which are dependent on the relevant kinds of false beliefs as irrelevant (and not just those which are also deliberatively irrational), we will move even further away from the deliberative idea. (This follows from the fact that

it is highly unlikely that all desires which possess this feature would extinguish in cognitive psychotherapy; some desires may simply be too deeply rooted).

To conclude: The "deliberative" version of the restricted rationality-oriented desire theory should not be accepted, i.e. it may well have nonderivative value for a person to have a certain desire fulfilled, even if this desire would not survive cognitive psychotherapy. Instead, we should accept the following claim: An underived desire is irrelevant if (and because) it is causally dependent on (maintained by) certain kinds of false beliefs (or "ignorances"), viz. on beliefs (etc.) that are (intuitively) about the relevant kinds of objects[90].

Rationality-oriented modifications based on the genetic theory

So, is it plausible to regard all "genetically irrational" desires as irrelevant? If not, should we (at least) accept the weaker claim that it is *ceteris paribus* better for a person to have his genetically rational desires fulfilled than to have his genetically irrational desires fulfilled?[91]

To get in a better position to answer these questions, let us first take a closer look at what the genetic theory is all about. On this theory, the difference between rational and irrational desires is a difference in causal history. A desire is rational if and only if (and because) it has been shaped "in the right way", and a desire is irrational if and only if (and because) it has been shaped "in the wrong way". This is how Elster (1983) formulates this idea:

> My suggestion is that we should evaluate the broad rationality of beliefs and desires by looking at the way in which they are shaped. A belief may be consistent and even true, a desire consistent and even conformable to morals - and yet we may hesitate to call them rational

---

[90]But it may still be objected: "*Why* should we regard all underived desires of this kind as irrelevant? How do you argue for the view that it is not good for a person to have such desires fulfilled?". Well, it seems that all I can do is to appeal to the intuitive plausibility of the idea.

[91]One may also ask whether it matters (in this context) in what way a desire is genetically irrational, e.g., what distorting mechanism that has generated it. Is it, for example, only certain kinds of genetically irrational desires (e.g., adaptive preferences) which should be regarded as irrelevant? It is not necessary to deal with these questions here, however.

> if they have been shaped by irrelevant causal factors, by a blind psychic causality operating 'behind the back' of the person. The stress here should be on 'irrelevant' and 'blind', not on causality as such. I am not arguing that beliefs and desires are made irrational by virtue of having a causal origin. All desires and beliefs have a (sufficient) causal origin, but some of them have the wrong sort of causal history and hence are irrational (pp 15-16).

On Elster's version of the genetic theory, a desire is irrational if and only if (and because) it has been "shaped by irrelevant causal factors". A desire is rational, on the other hand, if and only if (and because) it is *autonomous*. It is not easy to tell what the autonomy of a desire consists in, however. According to Elster, it "appears insuperably hard to say what it means for a desire to have been formed 'in the right way', i.e. not distorted by irrelevant causal processes" (ibid., p 20), and he admits that he can "offer no satisfactory definition of autonomy" (ibid., p 21). Instead, he characterizes it in a negative way; "autonomy will have to be understood as a mere residual, as what is left after we have eliminated the desires that have been shaped by one of the mechanisms on the short list for irrational preference-formation" (ibid., p 24)[92].

The distorting mechanisms to which Elster refers can be either cognitive or affective in character: desires may be irrational "due either to faulty cognitive processes or to undue influence from some affective drive" (ibid., p 24). An example of the former (cognitive, or "cold") kind of distorting mechanism is *preference change by framing*, which "occurs when the relative attractiveness of options changes when the choice situation is reframed in a way that rationally should make no difference" (ibid., p 25). An example of the latter (affective, motivational, or "hot") kind of distorting mechanism is *adaptive preference formation*. This is "the adjustment of wants to possibilities - not the deliberate adaptation favoured by character planners, but a causal process occurring non-consciously. Behind this adaptation there is the drive to reduce the tension or frustration that one feels in having wants that one cannot possibly satisfy" (ibid., p 25). Preferences shaped in this way

[92]It is worth noting that in the case of belief, it might be easier to characterize (or define) "genetic rationality" in positive terms, e.g., there is always the view that a belief is rational if it has been shaped by a process which is good at generating *true* beliefs. What makes the case of desire so problematic is (I think) the fact that there is no counterpart to "truth".

are called *adaptive preferences* (or "sour grapes"), and they are by far the most prominent kind of non-autonomous (irrational) preferences. Other important examples of heteronomous preference formation mentioned by Elster are "conformism, i.e. the adaptation of one's preferences to those of other people"[93], "sheer inertia", "counteradaptive preference formation" (which generates counteradaptive preferences like "the grass is always greener on the other side of the fence" or "forbidden fruit is sweet"), "anti-conformism", and "the obsession with novelty" (cf. ibid., p 22).

If we compare the genetic theory with the deliberative theory, we will see that the rationality conditions of the former theory are (most probably) "stronger" than the rationality conditions of the latter. It is highly likely that all autonomous desires would survive cognitive psychotherapy, but we can hardly assume that all heteronomous desires would extinguish in such a process. Suppose that my desire that there is peace in the world originated in a heteronomous way (e.g., I copied it from my peers), but that it is (so to speak) "maintained" in an autonomous way (the reason why I still have it is dependent on some "autonomous factor"). It is not likely that this desire would extinguish in cognitive psychotherapy[94]. In short, it seems plausible to assume that if a desire is "genetically rational", it is also "deliberatively rational", but not vice versa.

Another interesting difference between the two theories is this: It is easy to see how Elster's theory can (in principle) be used to classify desires and aversions as rational or irrational, but it is not clear whether it can also be used to determine whether it is rational or irrational for a person to desire a certain situation *to a certain degree*. Brandt's theory, on the other hand, is as good in this respect as it is to determine whether it is rational or irrational to desire a certain situation. The reason for this is that Brandt's theory is really a theory about what it is for a person's preference-ordering (as a whole) to be rational. Elster's theory is not "holistic" in this sense, however, and this means that he would have to adopt the following "atomistic" procedure in order to determine a person's rational preference-ordering: "Take a person's

---

[93]Cf. what Brandt says about artificial desires in note 86.

[94]What I regard as the strongest criticism against the genetic theory is connected to this point. The question is: "Is it really plausible to regard heteronomous desires of this type as irrational?"

actual preference-ordering, then eliminate every single outcome which is not an object of an autonomous desire or aversion, and what you are left with is a person's rational (i.e. autonomous) preference-ordering"[95].

Let us now return to the issue of relevance and ask whether it is plausible to regard all "genetically irrational" desires as irrelevant. And more generally, is it (in the present context of well-being) of any interest at all in what way our desires have been shaped?

Well, it is probably true that many heteronomous desires are best regarded as irrelevant, but this does not allow us to conclude that any of these desires irrelevant *in virtue of* having been shaped "in the wrong way". The idea that the relevance or irrelevance of a desire depends on how it has been formed is simply not valid, and should therefore be rejected. Instead, we should accept the more plausible view that desires which have the wrong kind of causal history will sometimes get (as a result of this) certain other defects, and it is in virtue of these defects that they are irrelevant. That is, we should regard the connection between genetical irrationality and irrelevance as contingent rather than necessary, and we should reject the idea that only autonomous desires are relevant. To see why it (in this context) doesn't really matter at all how a desire was shaped in the first place, consider a desire which was formed by adaptive preference formation, but which is maintained "in the right way". It is hardly plausible to regard such a desire as irrelevant.

To conclude, the idea that only autonomous desires can be relevant is a bad idea, and so is (for roughly the same reason) the idea that our heteronomous desires is less relevant than our autonomous desires[96].

---

[95]There are (of course) also a number of similarities between the two theories, e.g., they are both relativist conceptions of rational desire (they both allow for the possibility that it is rational for one person to desire a certain thing but irrational for another person to desire "the same thing"), and they both reject the Humean idea that only derived desires can be criticized on rational grounds.

[96]But that the idea should be rejected in the context of well-being does not imply that it should also be rejected in other contexts, e.g., in the context of rational action (or rational choice). For example, Elster may well be right in assuming that only autonomous desires are (fully) relevant in the context of rational action, i.e. that an action is not fully rational (rational "in the broad sense") unless it is rationalized by an autonomous desire (and a rational belief).

Rationality-oriented modifications based on the intrinsic theory

Assuming that there are such things as intrinsically irrational desires, is it plausible to regard all desires of this kind as irrelevant? And do we have a good reason to regard desires which are intrinsically rational (assuming that there are such things) as more relevant than desires which are not?

So, what is it for a desire to be intrinsically rational (or irrational)? Well, as a first approximation, we can say that a desire is intrinsically rational (irrational) if and only if it is rational (irrational) because of its content, i.e. no matter who has it. But what exactly does the intrinsic rationality (irrationality) of a desire consist in?

Let us first look at what it is for a desire to be intrinsically *rational* (or "rationally required", as Parfit puts it). In Parfit's (1984) terminology, a rationally required desire is a desire that provides good reasons[97] for acting, and he also claims that one is (in some sense) irrational if one does not have it (cf. p 118). I take this to mean that a desire is rationally required if it is rational to act *as if* one had it (even if one does not, in fact, have it). On this view, it is really actions and their reasons that are rationally required, rather than desires! Suppose, for example, that the desire to act morally (e.g., not to lie) is rationally required in this sense. What this means is that we all have a good (conclusive) "reason for acting morally", and that we are (therefore) rationally required to act morally. And "we have a reason for acting morally, even if we have no desire to do so" (ibid., p 121). This is what this part of the intrinsic theory melts down to: the claim that there are certain things which all of us have good reason to do (which is rational for everyone to do), no matter what we want and do not want. What is primarily required is that we act in a certain way, not that we have a certain desire[98]. The case of intrinsically *irrational* desires is (I believe) analogous. These desires do not provide good reasons for acting (cf. ibid., p 118), and

---

[97]That is, our desires *are* seldom our reasons, but they sometimes *provide us* with reasons. "In most cases, someone's reason for acting is one of the features of what he wants, or one of the facts that explains and justifies his desire. /.../ [M]y reason is not my desire but the respect in which my aim is [or better: appears to me as] *desirable* - worth desiring" (Parfit (1984), p 121).

[98]That is, if desires are rationally required at all, it is "by implication", i.e. because actions are (by definition) manifestations of desires.

every person is (therefore) rationally required to act *as if* he did not have it (even if he, in fact, happens to have it).

This suggests that a person's desire to act in a certain way is intrinsically rational if and only if he has an objective reason to act in this way. And if person's desire is a desire that a certain situation obtains (rather than a desire to act in a certain way), it is intrinsically rational if and only if he has an objective reason to (try to) bring the situation about, i.e. if it has objective value in Nagel's sense. This suggests that the question "How do we determine whether a certain desire is intrinsically rational, intrinsically irrational, or neither?" coincides with fundamental ethical questions like "How do we determine whether a certain situation is good or bad?" and "How do we determine what we have (objective) reason to do?". (It also suggests that many intrinsically rational desires are other-regarding, and that they must (for this reason) be regarded as irrelevant).

If we restrict our attention to now-for-now desires about one's own life, it seems plausible to assume that such a desire is intrinsically rational if and only if its object (some part of one's life) is worth desiring (if it has positive "desirability-value"[99]), and that it is intrinsically irrational if and only if its object is "in no respect worth desiring" or "worth avoiding" (if it has neutral or negative "desirability-value")[100]. (Here, it is important to note that the "desirability-value" of a situation, assuming that there is such a thing, is independent of what we want and do not want, and therefore, in a certain sense, objective. This does not imply that there are objective prudential values, however, i.e. that there are

---

[99]The term was suggested to me by Krister Bykvist.

[100]This is not the only way in which preferences about one's own life can be intrinsically irrational, however. According to Parfit, "[the] best examples [of such desires] can be found when we turn to our /.../ desires about possible pains and pleasures. Such desires are irrational if they discriminate between equally good pleasures, or equally bad pains, in an *arbitrary* way. It is irrational to care less about future pains because they will be felt either on a Tuesday, or more than a year in the future. /.../ In these cases the concern is not less because of some intrinsic difference in the object of concern. The concern is less because of a property which is purely positional, and which draws an arbitrary line. These are the patterns of concern that are, in the clearest way, irrational. These patterns of concern are imaginary. But they are cruder versions of patterns that are common. Many people care less about future pains, if they are further in the future. And it is often claimed that this is irrational" (ibid., pp 125-126). These preferences are not of a "now-for-now" type, however, and they are (therefore) of no interest in the present context.

situations which are nonderivatively good or bad for us, regardless of what our desires and aversions are. Cf. below).

Let us now return to the issue of relevance, and ask whether we should regard all intrinsically irrational now-for-now desires about one's own life as irrelevant. Can it ever have nonderivative value for a person to have a desire about his own life fulfilled if the object of this desire has negative desirability-value? Well, this seems to depend entirely on what kind of desirability-value we have in mind. For example, if the object is "desirability-bad" in the aesthetic sense, or if it is morally bad, or instrumentally bad-period, it may well be good for the desiring subject to have the desire fulfilled. But if the object of a desire is desirability-bad in the prudential sense, it is (obviously) not good for the subject to have this desire fulfilled.

On my view, prudential desirability-value is the only type of desirability-value that is of interest here, i.e. the only restriction claim that we have any reason to accept in this context is this: A desire can only be relevant if it is a desire for something that does not have negative nonderivative desirability-value for the subject. And in a similar way, we should regard the desires whose objects have positive prudential desirability-value as more relevant than the desires whose objects have neutral or negative prudential desirability-value[101]. It is worth noting that both these claims are perfectly consistent with (D1).

The big question here is of course whether there are any desires which are intrinsically rational (or irrational) in this way. Now, this is exactly what one of the weaker forms of "the objective list theory" claims, i.e. the question is really whether this theory is plausible or not (cf. chapter 7, pp 362-366).

Conclusion

To sum up, it seems that we have reason to accept the following rationality-oriented restrictions of the Success Theory:

(i) All derived desires that are not (in a strict sense) derivable from the truth about "their objects" (and more fundamental desires) are irrelevant. That is, a derived intrinsic desire is irrelevant if there is no way

[101]That is, if X has positive prudential desirability-value while Y has not, then it is *ceteris paribus* nonderivatively better for a person to have the (intrinsic) desire that X fulfilled than to have the (intrinsic) desire that Y fulfilled.

in which the subject could have derived it correctly from the whole truth about "its object" and his more fundamental desires, e.g., if it is (so to speak) logically dependent on mistakes of fact (cf. p 241 above).

(ii) An underived desire is irrelevant if (and because) it is causally dependent on (maintained by) certain kinds of false beliefs or "ignorances", viz. on beliefs (etc.) that are (intuitively) about the relevant objects. The idea is (roughly) that there must be some conceptual connection between the object of the desire and the object of the "irrelevance-making" belief, i.e. that the propositional content of the desire and the propositional contents of the relevant beliefs must, to some extent, "coincide"(cf. p 253)[102].

(iii) On the assumption that there are situations that are "worth avoiding" (in the prudential sense): A desire can only be relevant if it is a desire for a situation that is *not* of this type (cf. p 259).

We should also accept the following rationality-oriented claim concerning relative weights: On the assumption that there are situations that are worth desiring (in the prudential sense): Desires whose objects are of this type are more relevant than desires whose objects are not of this type, e.g., desires for situations that are (in the prudential sense) "in no respect worth desiring" (cf. p 259).

This ends the section on the different rationality-oriented ways in which the Success Theory may (and should) be modified. So, are these all the modifications we need, or are any further modifications necessary? Is the rationality-oriented Success Theory we have arrived at the most plausible version of the actual desire theory, or is it necessary to modify this theory too, i.e. in a way that is neither object-oriented nor rationality-oriented?

---

[102]Claims (i) and (ii) are both "based on" the distinction between derived and underived desires. So what if this distinction cannot be upheld? Well, then it seems that we should simply drop (i) and conceive of (ii) as applicable to all desires, i.e. not just underived ones.

5.2.3. Some other possible modifications of the (actual) desire theory

In this section, I will take a closer look at two types of possible modifications, viz. (1) modifications which "appeal to the agent's values, or ideals, or to his moral beliefs" (cf. Parfit (1984), p 119), or, we should add, to his higher-order desires, and (2) "knowledge-oriented" modifications. The modifications of the first type can all be regarded as versions of the following general claim: "How good it is for a person to have a certain desire fulfilled depends (in part) on how compatible it is with his own higher-order desires, evaluations, ideals, or moral beliefs. Or more specifically, those of a person's desires that are sanctioned by his higher-order desires (etc.) should be regarded as more relevant than those of his desires that are not". The knowledge-oriented modifications are of two types, viz. (a) the idea that we should give less or no weight to those of a person's desires of which he is unaware, and (b) the idea that it is *ceteris paribus* better for a person to have a desire fulfilled if he is aware of the occurrence of its object[103].

The appeal to evaluations and higher-order desires

Modifications of the first type come in several forms, e.g., it might be claimed that we should "give no weight to the desires that someone wishes that he did not have" (cf. ibid., p 119), or that we should "only /.../ pay attention to the preferences we welcome and accept", the preferences "we gladly and willingly acknowledge" (cf. Tännsjö (forthcoming), p 83). It might also be claimed that we should give more weight to the desires with which the subject identifies, or acknowledges as (in some important sense) his own (cf. Sumner (1996)), or that "our

[103]Apart from the content-oriented and rationality-oriented modifications above, these are the only types of modifications to which I will pay any attention. This means that there are many possible modifications which will be ignored, e.g., the idea that desires with certain kinds of causal histories are more relevant than desires with other kinds of histories, or the idea that desires that are in harmony with human nature are more relevant than desires that are not, or the idea that we should give more weight to those desires that serve our needs than to those desires that are in conflict with our needs. The first of these three modifications is intuitively implausible and the last two are nothing but clumsy reformulations of the rationality-oriented modifications on p 259, so it should hardly surprise us that none of these possible modifications can (as far as I know) be found in the literature.

lives are subjectively good if we succeed in satisfying our most important wants *in accordance with appropriate ideals*", i.e. that it is a "requirement of good lives /.../ that the satisfaction of /.../ [our] most important /.../ wants must occur *in accordance with ideals we correctly value*" (Kekes (1988), p 19, my italics)[104].

The fact that there are so many possible versions of the general claim suggests that it is not very fruitful to ask whether this claim is plausible or not. Instead, we need to be more specific: we have to identify those ("major") versions of the claim that are intuitively interesting, and then ask (for each of these claims) whether it is plausible.

On my view, there are at least two such "sub-claims" that are worth investigating, viz. (i) the idea that we should give less or no weight to those of a person's desires and aversions that are not sanctioned by his higher-order desires, or by those of his evaluations (moral or not) whose objects are (explicitly or implicitly) desires or aversions, and (ii) the idea that we should give less or no weight to those of a person's desires that are not sanctioned by those of his evaluations (moral or not[105]) whose objects are *not* desires or aversions[106].

---

[104]It is worth noticing that this restriction claim consists of two distinctive parts, viz. (i) the idea that a person's desire is not relevant unless it (and/or its satisfaction) is in accordance with her ideals, and (ii) the idea that the ideals in question must be appropriate. The latter idea is not as central as the first, but there will be some discussion of it anyhow.

We might also add that this claim is but one of several components of Kekes' conception of the good life. The other components of his desire theory are: (a) The idea that "desires to do" are (somehow) more important than desires with other kinds of objects, e.g., as when he says that "[t]he central idea I propose to advance and justify is that good lives depend on *doing* what we want" (ibid., p 18, my italics); (b) the idea that a want can only be relevant if it is both rational and moral, i.e. that our desires are not relevant if their objects (what we want to do) are "foolish, inconsistent, or immoral things" (ibid., p 18); and (c) the idea that a desire can only be relevant if it "matters deeply", if it is not "transitory and superficial", and that the more important a desire is (for the desiring subject), the better it is for him to have it satisfied (that is, his desire theory is intensity-oriented). However, Kekes' theory also contains (d) the objectivist idea there are certain things (e.g., intimate personal relationships) that a good life must include, i.e. it is not a "pure" desire theory.

[105]In this context, it does not really matter whether a person's values and ideals are "moral" or not, and "those of a person's desires that are sanctioned by his moral beliefs" is (for this reason) not an interesting category.

[106]The distinction between evaluations of this type and the evaluations referred to in (i) corresponds to the distinction between first-order desires and higher-order desires, and we might therefore (somewhat inappropriately) refer to them as "first-order evaluations" and "higher-order evaluations", where "pain is a bad

It should also be added that both (i) and (ii) come in two forms, i.e. they either take the form of restriction claims, where all "non-sanctioned" desires are regarded as totally irrelevant, or of claims concerning relative weights, where "non-sanctioned" desires are given some weight, but not as much weight as those desires that are (in the relevant sense) "sanctioned".

So, are (i) or (ii), in any of their forms, plausible? Well, to be able to answer this question, there are two things we need to know, viz. what it is for a person's desire to be sanctioned by his higher-order desires (or evaluations), and what it is for a person's desire to be sanctioned by his "first-order evaluations".

In my terminology, a person's desire is *minimally sanctioned* by his higher-order desires if and only if he does not have an aversion against having it, and it is *maximally sanctioned* by his higher-order desires if and only if he (positively) wants to have it. A person's first-order desire is *in conflict with* his higher-order desires if and only if it is not minimally sanctioned by them, i.e. if he prefers not to have it. In a similar way, a person's desire is minimally sanctioned by his "higher-order evaluations" if and only if he does not regard its occurrence as bad, and it is maximally sanctioned by his "higher-order evaluations" if and only if he regards its occurrence as good (and so on).

A person's (first-order) desire is *minimally sanctioned* by his "first-order evaluations", on the other hand, if and only if he does not evaluate its object (i.e. some situation) in a negative way, and it is *maximally*

---

thing" is an example of a "first-order evaluation", and where "it is good to have benevolent desires" is an example of a "second-order evaluation".

So, what about the idea that we should give less or no weight to those of a person's desires that are not sanctioned by his *ideals*? Isn't this an "intuitively interesting" idea? Well, I think it is, but on my view, it doesn't really contain anything that isn't already contained in (i) and (ii). What we call "ideals" are most often mixtures of desires and evaluations: Ideals normally include certain kinds of evaluations - viz. evaluations of the form "an ideal (good, perfect) X (e.g., a good society, an ideal partner, or a perfect family) is an X with the properties P, Q, and R" - but full-fledged ideals are more than just evaluations; they are also *goals*. (An evaluative ideal is not a full-fledged ideal of *mine*, it is not *my* ideal, unless I am (in some way or other) committed to its realization). It is also worth mentioning that some of the desires of which the relevant ideals "consist" are higher-order desires. For example, the ideals which tend to be of most interest in this context, viz. "personal" ideals concerning what kind of person (man, woman, etc.) one wants to be, normally involve some reference to what kind of preferences one wishes to have (this follows from the fact that the preferences one has is in part constitutive of what kind of person one is).

sanctioned by his "first-order evaluations" if and only if he regards the object of the desire as good. A person's desire is *in conflict with* his "first-order evaluations" if and only if he conceives of its object as bad.

So, now that we know what it is for a desire to be sanctioned by his higher-order desires and/or his evaluations: Are any of the modifications suggested above plausible? Well, let us first note that it would be highly implausible to require that a person's desire is *maximally* sanctioned by his higher-order desires and/or evaluations in order to count as relevant. This suggests that we may, in the "higher-order case", restrict our attention to the following questions: (i:a) Should we regard those of a person's desires which are minimally sanctioned by his higher-order desires (or evaluations) as more relevant than those of his desires which are in conflict with his higher-order desires (etc.)? (i:b) If the answer is yes: Should we also regard all first-order desires that are in conflict with higher-order desires (or evaluations) as totally irrelevant? (i:c) Is it plausible to regard those of a person's desires which are maximally sanctioned by his higher-order desires (or evaluations) as more relevant than those of his desires which are only minimally sanctioned by his higher-order desires (etc.)?[107] (To complicate the issue further, we may also ask (i:d) whether it matters whether the relevant (i.e. "sanctioning" or "non-sanctioning") higher-order desires (etc.) are intrinsic or instrumental, and (i:e) whether it makes any difference whether they are rational or not; and if it does, whether it matters in which way they are rational[108]).

Let us first look at what seems to be the most central question here, viz. (i:b). Suppose that Alice has an intrinsic (first-order) desire to smoke, but that she also wishes that she did not have this desire. Suppose also (for the sake of argument) that this second-order desire is both rational[109] and intrinsic, i.e. that it is relevant. Does this allow the

---

[107]And similarly in the "first-order evaluation case", i.e. (ii) can, in a similar way, be divided into three sub-questions.

[108]In the "first-order evaluation case", the corresponding questions are: (ii:d) Does it matter what kinds of evaluations we have in mind, e.g., whether the "sanctioning" (or "non-sanctioning") evaluation is intrinsic or instrumental, and whether it is aesthetic, moral or prudential? (i:e) Does it make any difference whether the relevant evaluations are "appropriate" ("correct") or not (as Kekes claims)?

[109]The assumption that this second-order desire is rational might be somewhat problematic, however, viz. because it may well be incompatible with the necessary assumption that the first-order desire (i.e. the desire to smoke) is (in the

desire theorist to "conclude" that it does *not* have nonderivative value for Alice to have her first-order desire fulfilled?

Well, it may seem so, viz. for the following "reason": "If we want to determine which of a person's desires that are relevant and which of his desires that are not, why don't we just appeal to the Sovereign Desiring Subject, and let him decide? After all, this is the restriction claim which is most faithful to the subjectivist spirit of the desire theory". This is a bad argument, however. First, if we really want to be faithful to the idea of the Sovereign Subject, we should probably not restrict at all. Second, we have already admitted (in the spirit of the desire theory) that the second-order desire should count as relevant, so why should we count it again, by using it as a "criterion of relevance"?

On my view, this is the strongest objection to the idea that only desires that are sanctioned by higher-order desires should count as relevant: There is really no reason why we should (in this way) give special weight to those desires which have other desires as objects. As has already been suggested (on p 213): If we understand the notions of desire and strength in the relevant way, i.e. in terms of preference orderings, we can see that our higher-order desires are already (in a certain way) "incorporated" into our first-order preferences (it is highly likely that a person's "first-order preference ordering" is affected by his higher-order desires in a way that the felt intensities of his first-order desires are not). And if we are aware of this fact, it is hard to see why we should also (on top of this) use our higher-order desires as a "criterion of relevance". There is simply something arbitrary about giving our higher-order desires that much weight, and my tentative conclusion is therefore that we should give a negative answer to (i:b)[110].

---

relevant sense) rational.

[110]But suppose I am wrong here, i.e. that Alice's desire to smoke should (in fact) be regarded as irrelevant. We would then have to ask ourselves whether it would make any difference if her higher-order desire to get rid of the desire to smoke were instrumental rather than intrinsic, or if it were irrational rather than rational. On my view, it wouldn't really change anything if the second-order desire were (instead) instrumental, but it would make a difference if it were irrational, e.g., it might be argued that irrational higher-order desires should not count, or more specifically, that if a certain desire is in conflict with an irrational higher-order desire, then this does not make the desire irrelevant, and if a certain desire is sanctioned by an irrational higher-order desire, then this does not "add" to its relevance at all.

Now, it seems that this argument can also be used to show that (i:a), (i:c), and (ii:a)-(ii:c) should be answered in the negative. That is, my conclusion is that the rationality-oriented Success Theory we have arrived at is *not* in need of any modifications of this type. So, let us now see whether it needs to be modified in a "knowledge-oriented" way.

Knowledge-oriented modifications

To repeat, there are two kinds of knowledge-oriented modifications of the desire theory, viz. (a) the idea that we should give less or no weight to those of a person's desires of which he is not aware, and (b) the idea that it is *ceteris paribus* better for a person to have a desire fulfilled if he is aware of the occurrence of its object.

As usual, both (a) and (b) come in two forms, i.e. they either take the form of restriction claims, e.g., where all desires of which the subject is unaware are regarded as totally irrelevant, or of claims concerning relative weights, e.g., where those desires of which the subject is aware are given more weight than those desires of which he is not aware.

So, are (a) or (b), in any of their forms (in particular, the restriction claims), plausible? Well, to be able to answer this question, we first need to know what it is for a person to be *aware* of the fact that a certain situation obtains (e.g., that he himself has a certain desire). So, what do I mean by "awareness" in this context? Well, suffice it to say that the notion of awareness I have in a mind is a broad one: A person is (of course) aware of X if he knows that X is the case, but also if he has a true belief that X is the case. Moreover, we must (I think) not assume that awareness is necessarily propositional, i.e. we should allow for the possibility that there is a kind of awareness which has no belief-component. However, it seems necessary that the object of such non-propositional awareness is (somehow) closely related to some proposition or other (it is, after all, "*awareness of facts*" that we are interested in).

So, now that we some idea of what is meant by "awareness" in this context: Are any of the modifications suggested above plausible? In particular, is any of the following two "restriction claims"[111] plausible:

(a) It can not be nonderivatively good for a person to have a desire fulfilled unless he aware of having the desire. Or more specifically, if a

[111]Again, it may be pointed out that it is only (a) that is a restriction claim in the proper sense. Cf. note 44 on p 220.

person has (at *t*) an intrinsic desire that X, then it can only have non-derivative value for P-at-*t* to have the desire fulfilled (at *t*) if he is (at *t*) aware of the fact that he desires that X[112].

(b) It can not be nonderivatively good for a person to have a desire fulfilled unless he is aware of the occurrence of its object. Or alternatively put: On the notion of fulfilment which has most "moral and rational significance", a desire is not fulfilled unless the subject is aware of the occurrence of the object (cf. p 162 above)[113].

Now, it seems fruitful to deal with both these questions at the same time, viz. as follows. Let us first give a schematic representation of the possible epistemic situations of a desiring subject. If we let "P" refer to the desiring subject, "D" to some desire, and "X" to the object of this desire, we can represent the different possibilities in the following simple way:

	P is aware of X	P is not aware of X
P is aware of D	A	B
P is not aware of D	C	D

We can now formulate the problem as follows. If we assume that P has an intrinsic now-for-now desire that X, and that this desire is relevant in all other respects (e.g., it is about his own life, it is rational in the relevant sense, and so on): Under what epistemic circumstances does it have nonderivative value for P to have this desire fulfilled?

It is obvious that it is (on all plausible versions of the desire theory) good for P to have the desire fulfilled in situation (A), i.e. if he is aware both of the fact that he has the desire and the fact that the object obtains[114]. It is (I think) also obvious that it is *not* good for P to have the

---

[112]That is, (i) the important thing is that the person is aware of *what* he desires, i.e. it is not necessary that he knows how strong his desires are; and (ii) if *t* is the time when the desire is fulfilled, it is not necessary that P knows *before t* that he has the desire (assuming that it existed before it was fulfilled).

[113]This is but a special case of the more general idea that a person's well-being can not be directly affected by things he doesn't know anything about, that if a certain situation does not (in any way) "enter" or "affect" a person's experience (if the person is not aware of its occurrence), it can not have nonderivative value for the person that the situation obtains (cf. Sumner (1996), pp 124-128).

[114]This suggests (but I am not sure that it implies) that he is also aware of the fact that the desire is fulfilled, i.e. of the fact that the occurrence of X constitutes the fulfilment of the desire.

desire fulfilled in situation (D); cf. argument (iii) on p 214 above.

(C) is a strange kind of situation; it is (I think) doubtful whether it is really possible to be aware of a desired situation without being aware of it *qua* desired (what notions of desire and awareness does this presuppose?). But if we assume that it is possible: Is it good for P to have the desire fulfilled in this type of epistemic situation? Well, I think not, i.e. I tend to accept restriction claim (a) on pp 266-267.

So, what about cases of type (B)? If P is aware of having a desire that X, and if this desire is, unknown to P, fulfilled[115], is this good or not-good for P? To find an answer to this question, let us look at some examples. Suppose that Bert has a strong intrinsic desire to be loved by Alice, and that he is aware of this fact. Suppose also that Alice actually loves Bert, but that he has no awareness whatsoever of this fact. On my (tentative) view, it is doubtful whether it is, in this case, non-derivatively good for Bert to have the desire fulfilled. Now, consider another example: Suppose that Bert has a strong intrinsic desire not to be deceived and betrayed by Alice, and that he is aware of this fact. Suppose also that he is (in fact) deceived and betrayed by Alice, but that he is not aware of this. Is it, in this case, bad or not-bad for Bert to have the aversion fulfilled? Well, personally, I tend to regard it as bad-for-Bert[116]. But is this really compatible with the idea that it is (in the example above) not-good for Bert to be loved by Alice? Well, I think it is; either because there is (in this context) an asymmetry between desire and aversion, between good and bad, or because being deceived and betrayed have negative prudential desirability-value while being loved does not have any positive prudential desirability-value.

My tentative conclusion is that we should accept the following "awareness-oriented" modifications of the desire theory:

(a) If a person has an intrinsic desire that X (or aversion to Y), then it can only have nonderivative value (positive or negative) for P to have the desire (aversion) fulfilled if he is aware of the fact that he desires that X (has an aversion to Y)[117].

---

[115]In the traditional sense, that is.

[116]Is it hard to say whether this is a plausible idea, though. Some people accept the idea that a person's well-being can be directly affected by things he doesn't know anything about, and other people reject it, and it is hard to see how the issue could be settled in a rational way. As far as I can see, this is one of the truly intractable disagreements which prevail in this field.

[117]That is, awareness of this kind is necessary for relevance. It is far from

(b) It can not be nonderivatively *good* for a person to have a desire fulfilled unless he is aware of the occurrence of its object. However, it can (at times) be nonderivatively *bad* for a person to have an aversion fulfilled, even if he is unaware of the occurrence of its object, viz. if this object has negative prudential desirability-value.

At this point, it is worth noting that condition (a) gives rise to at least four questions:

(i) The fact that it makes perfectly good sense to say things like "Now I realize that I have always had this desire, a month ago I didn't fully realize it, about a year ago I began to realize it, and before that I had no idea whatsoever that I had it" suggests that a person can be more or less aware of the fact that he has a certain desire. That is, it seems that the difference between the desires of which we are aware and the desires of which we are not aware is really a matter of degree. So the question arises: Does this mean that we have to accept the idea that the more aware a person is of the fact that he has a certain (intrinsic) desire, the nonderivatively better it is for him to have it fulfilled, e.g., should we regard the desires of which we are "semi-aware" as "semi-relevant"? On my view, we should not. Instead, we should draw an "epistemological line" somewhere, and regard all the desires below the line as irrelevant. That is, if the degree to which a person realizes that he has a certain intrinsic desire is not high enough, then it is not good for him to have the desire fulfilled. This means that even though awareness is a matter of degree, *relevance* will (in this area) still be an all-or-nothing matter.

(ii) There are several different ways in which we can be aware of our desires, and in the special case of knowledge, there are several ways in which we can gain knowledge of our desires, not just about what we want but also about how much we want it. In some cases, we have *non-inferential knowledge* of what we want[118]: we know what we want directly (through "introspection"), without having to infer this from observations of what we do and do not do (or what we experience)[119].

---

sufficient, however.

[118]Do we also, at times, have non-inferential knowledge of *how much* we want something? Perhaps, but it is (I think) not a very common thing. In any case, it is important to notice that the fact that we may find out *what* we want through introspection does not (in any way) guarantee that we, in this way, can discover *how much* we want this something.

[119]So, how is it possible to have such non-inferential knowledge about one's own

In other cases, we have *inferential knowledge* of our desires, we gain knowledge about what we want and how much we want it by inference from certain kinds of observations, viz. from observations of what we do and do not do (what we try to do and not to do), and from observations of what we perceive, think, feel, dwell on, focus on, and so on[120].

Now, the fact that our knowledge about our desires are of different kinds gives rise to the following question: Does it matter (in this context) what kind of knowledge the desiring subject has about what he wants (and how much he wants it), e.g., whether it is inferential or noninferential? For example, if he knows that he has an intrinsic desire that X obtains, is it better for him to have the desire fulfilled if his knowledge about it is noninferential than if his knowledge is "merely" inferential[121]? I think not. As long as the subject knows that he has a certain desire, it does not really matter what kind of knowledge he has of it (and neither does it matter whether he knows that he has the desire or merely has a true belief that he has the desire).

---

desires? This is (I think) a difficult question, especially for a functional conception of desire: How can we know directly that we have a certain disposition?

[120]In this context, it is not really necessary to discuss exactly how knowledge of our desires can be inferred from knowledge of our actions and experiences. Suffice it to say that such inferences are often difficult to make, something which is connected to the fact that there are a number of conditions that have to be satisfied for such inferences to be possible. For example, it seems that reliable inferences from the fact that a person P performs an action A to the alleged fact that P has a desire D (with a certain strength) are not really possible unless the following conditions are met: (a) We have some knowledge about *P's beliefs*. More specifically, it can not just be assumed that P believes that A is a possible way to satisfy D, it can also be assumed that P believes that A is the best possible way open to her (or at least a comparatively good way) to satisfy D. And to know this, we have to know what actions P believes are open to him (where this can, most probably, not be determined in any independent way, i.e. without making any assumptions about what P wants). (b) We do not just know what options P *believes* he has, we also know what options are *actually* open to P, or (in Jeffrey's terms) "what propositions it is in his power to make true". That is, we can not infer what P wants from what he does (and does not do) unless we have some knowledge about what he can and cannot do, e.g., unless we know something about what abilities and opportunities he has, what he dares to do and does not dare to do, and whether he is akratic ("weak-willed") or not (in the relevant respects). (c) We know whether P is a normal person (given her culture, and so on) or not. Most importantly, we know whether or not his needs can be regarded as "normal human needs".

[121]That is, if two intrinsic and known desires are of the same strength, and if one of them is known "noninferentially" while the other is known "inferentially", is it better for the desiring subject to have the former desire fulfilled?

(iii) Condition (a) on p 268 says that it can not be nonderivatively good for a person to have the (intrinsic) desire that X fulfilled unless he is aware of the fact that he desires that X. So the question arises: Must he also be aware of the fact that the desire in question is intrinsic, i.e. that he desires that X *as an end*? Frankly, I don't know.

(iv) What would (and should) a desire theorist say about the following type of case: Suppose that a person P has this instrumental desire, but that he falsely believes that the desire is intrinsic[122]. Is it, in a case like this, nonderivatively good for P to have the desire fulfilled? I think not; the desire is (after all) not intrinsic.

To sum up, it seems that the rationality-oriented Success Theory we have arrived at is in need of some further modifications after all, viz. (a) and (b) on pp 268-269 above. That is, the most plausible version of the actual desire theory is a Success Theory that is both rationality-oriented and "awareness"-oriented. This theory can be characterized as follows:

My conclusion: The most plausible version of the actual desire theory

First, the most plausible version of the desire theory is a restricted theory: it claims that only some kinds of intrinsic now-for-now desires should be regarded as relevant. More specifically, if a person P has an intrinsic (etc.) desire that X, the theory claims that it is nonderivatively good for P to have the desire fulfilled (in the traditional sense) if and only if the following conditions are satisfied:

(i) X is a part of P's life.

(ii) If the desire is derived: It is derivable from the whole truth about "its object" (and more fundamental intrinsic desires).

---

[122]But can such a belief really be mistaken? Well, it seems that if someone believes that he desires that X, then it is very unlikely that his belief is false. (The mistakes we make about *what* we do and do not desire are normally of the following kinds: We have desires that we not believe (know) that we have, or we have desires that we believe that we do not have). But a person may well be mistaken about *how much* he wants something, and most importantly in this context, about what he does and does not desire *intrinsically*. (A psychologically interesting type of mistake is this: P correctly believes that he desires that X, but he does not know that he also desires that not-X (i.e. that he is ambivalent), and neither does he know that this unknown desire that not-X is in fact stronger than the desire that X).

(iii) If the desire is underived: It is not causally dependent on (maintained by) certain kinds of false beliefs or "ignorances", viz. on beliefs (etc.) whose propositional contents stand in a close enough conceptual relation to the propositional content of the desire.

(iv) X is not a situation that is "worth avoiding" (in the prudential sense).

(v) P is aware of the fact that X is (in fact) desired by him.

(vi) P is aware of the fact that X obtains. (But with the following proviso: It may sometimes be bad for a person to have an aversion fulfilled, even if he is unaware of the occurrence of its object, viz. if this object has negative prudential desirability-value).

Next, the theory also makes certain claims about how we should determine which of two relevant desires that it is better for the desiring subject to have fulfilled. The fundamental idea is of course that relevance is a function of strength, but there is one possible exception to this rule, viz.

(vii) the idea that desires for situations that are worth desiring (in the prudential sense) are more relevant than desires whose objects are not (in this sense) worth desiring.

(iv) and (vii) are both based on the "objectivist" assumption that there are situations that are (in the prudential sense) worth desiring, or in no respect worth desiring, or worth avoiding. This means that if this assumption proves to be false, the most plausible version of the desire theory is a restricted intensity-oriented theory which can be summarized in points (i)-(iii), (v), and (vi).

It is now time to move on to the second question on p 188, and ask whether this theory is a plausible theory of prudential value.

5.3. Is any version of the actual desire theory plausible?

If any version of the actual desire theory is plausible, it is (of course) the most plausible version, i.e. the rationality-oriented (and awareness-oriented) Success Theory that was presented above. So, is this theory a plausible theory of prudential value? Is the best theory of prudential

value a kind of desire theory at all?[123]

We have already seen that there are good reasons for answering these questions in the affirmative. To repeat, the argumentation for our modified Success Theory consists of two parts, viz. (i) arguments which purport to show (D1) is valid, i.e. that we have to accept some kind of desire theory (cf. pp 193-195), and (ii) arguments which purport to show that this particular theory is the most plausible version of the desire theory, i.e. that *if* (D1) is valid, then this is the version of the theory we should accept (cf. pp 205-214 and 223-272).

The fact that there are good reasons for accepting the modified Success Theory does not mean that we should accept it, however; there may (after all) be even better reasons for rejecting it. So let us now look at the arguments which have been (or can be) directed against the theory, to see whether these counter-arguments (or "counter-claims") are valid; and to the extent that they are, whether they have more weight than the pro-arguments. Here are the arguments that can be given for the view that the most plausible version of the desire theory is an implausible theory of prudential value:

(1) Counter-arguments of the first type purport to show that *if* you accept the most plausible version of the actual desire theory (or some similar type of restricted rationality-oriented desire theory), then you "must", in order to be "consistent", reject (D1). That is, these arguments are not directed against (D1) as such, but against (D1) *qua* component in a certain type of rationality-oriented (actual) desire theory, and they can therefore be viewed as "immanent criticisms" (they start from the opponent's own beliefs, and then try to use these beliefs to defeat him). There are two possible arguments of this type:

(a) "The fact that you have accepted the restriction claim that an actual desire cannot be relevant unless it is (in a certain sense) informed shows that you have realized that information matters. So why don't you also accept the corresponding "idealization claim", viz. the idea that we should also give some weight to what a person would want if he had (in fact) been fully informed? Isn't it rather arbitrary (for a rationality-oriented desire theorist) to exclude people's hypothetical desires in this way?"

---

[123]The most central question here is (of course) whether (D1) is a plausible claim, but there is (as we will see) more to the issue than this.

(b) "You believe that a desire cannot be relevant if it is a desire for something that is (in the prudential sense) worth avoiding, and you also believe that desires whose objects are (in this sense) worth desiring is more relevant than desires with other kinds of objects. Now, this view implies that there are situations which can be (in the relevant sense) worth desiring, even if they are not (in fact) desired. So why don't you admit that it can be good for us that such situations obtain, no matter what our actual desires are?"

Both these arguments are (of course) bad. The beliefs referred to are all compatible with (D1), not just in the literal sense but also "in spirit", and our actual desire theorist is (for this reason) not in any way inconsistent (he may, however, as we will soon see, be forced to reject (D1) for other reasons). Moreover, it is not just that he will reject the views that the critic thinks he should accept, it is also likely that he will reject them just because they are incompatible with (D1)[124].

(2) Arguments (or "counter-claims") of the second type are directed against (D1) as such. That is, *qua* arguments against the satisfaction interpretation, they purport to show that there are other things besides (actual) desire-fulfilment and aversion-fulfilment that have non-derivative value for a person, and *qua* arguments against the object interpretation, they purport to show that a situation may well have nonderivative value for a person even if it is not desired (i.e. that it is not the case that all prudentially good situations are good in virtue of being desired)[125]. There are at least three arguments of this type:

(a) The first argument appeals to the idea that if a person would intrinsically desire that X if he were fully rational (e.g., if he knew the relevant facts and were thinking clearly), then it is nonderivatively good for this person (as he is now) that X obtains.

(b) The second argument is based on a different idea, viz. the objectivist idea that there are certain things which are nonderivatively good

---

[124]Once again, we can see that "one person's modus tollens is another person's modus ponens" (this saying is attributed to Wlodek Rabinowicz).

[125]On my view, it is more appropriate to conceive of these "arguments" as "counter-claims" rather than as counter-arguments in the proper sense. The reason for this is that most of the "arguments" I have in mind have the following structure: They first make some key claim, and they then point out that (D1) is inconsistent with this claim, and that it should therefore be rejected. They do not really purport to show us *why* we should accept the key claim, however; instead, we are simply told *that* we should accept it.

and bad for us, regardless of what we want and do not want, and regardless of what we would want if we were rational (or informed). According to this line of criticism, the desire theory is defective because it fails to recognize that there are certain objective prudential values.

(c) The third argument is (roughly) a more radical version of the second. The second type of criticism is directed specifically against (D1), and it is therefore consistent with the idea that it is often nonderivatively good for us to have our desires fulfilled (it is even consistent with (UD2)). The third line of criticism, on the other hand, purports to show that the desire theory is not even "partly right", i.e. that there is no truth at all in the theory. *Qua* attack on the object interpretation of the theory, the argument purports to show that nothing can have nonderivative value for person in virtue of being desired. If it is good for a person P that a desired situation obtains, the reason for this is either that it has objective prudential value or that it gives P pleasure: it is never good for P *because* he desires it. If we, instead, view the argument as an attack on the satisfaction interpretation, it purports to show that even though it is sometimes good for us to get what we want, it is never the desire-fulfilment in itself that is good for us[126].

Let us now take a closer look at these three types of objections, to see if any of them are valid.

Is there any truth in the idea that hypothetical desires should count?

According to the *idealized versions* of the desire theory, it is the (hypothetical) preferences that we would have under certain ideal conditions that should be regarded as relevant. The preferences that it is good for a person to have "satisfied" are (on this view) not his actual preferences, but the preferences he would have had if he were fully rational, fully informed, free of neuroses, assessing the matter "in a cool hour", or the like. Or more precisely: A situation X is nonderivatively good for a person P (as he is now) if and only if (and because) P would

[126]Or more precisely: Suppose that P desires that X and that X obtains. On the view under consideration, it is the mere fact that the desire is fulfilled (i.e. the conjunction of the fact that P desires that X and the fact that X obtains) that cannot have nonderivative value for P (i.e. X itself may well be nonderivatively good for P).

want X to obtain if he were fully rational[127] (I assume that all idealized desire theories are rationality-oriented).

So, is this a plausible view? Should we regard a person's "hypothetical rational desires" as relevant? Suppose that a person would desire that X obtains if he were fully rational: Does this imply that is it good for him, *as he is now*, that X obtains, even if he does not actually desire that X? I think not, and for the following reasons:

Suppose that the actual Bert is an irrational, uninformed, confused, and neurotic person. Surely it cannot be good for this actual Bert to get something to which he now has an aversion, just because he would want it if he were an entirely different person[128]. Consider the following example: Suppose that Bert has this identical twin brother who is fully rational (in fact, he is one of the few persons on this planet who has undergone cognitive psychotherapy). Suppose also that if Bert were rational, he would be just like his brother, i.e. he would have exactly the same desires as the brother now has[129]. In this case, an idealized desire theorist would claim that "what is good for Bert's brother is also good for Bert"[130]. This is not a plausible view, however. Bert is not his brother, and his actual desires may be very different from his brother's actual desires. The fact that the brother wants to live a certain kind of life does not imply that it would be good for Bert to live the same kind of life: If Bert's global desires are very different from his

---

[127]The reason why I have (in this way) taken the object interpretation for granted is that satisfaction interpretations of idealized theories do (on my view) not really make sense. It does not make sense to attribute prudential value to "the circumstance that our hypothetical (intrinsic) desires and preferences be satisfied", and the reason for this is that non-existing desires cannot be fulfilled (in any literal sense). Suppose that a person P has no (actual) desire that X obtains, but that he would desire that X if he were fully rational. Now, if X (in fact) obtains, this is hardly a case of desire-fulfilment, since there is no desire there to be fulfilled. Or alternatively put, it hardly makes sense to attribute value-for-P to the conjunction "P would desire that X if he were rational *and* X obtains". (But it makes perfectly good sense to say that if P would desire that X, then it *would* be good for P to have it fulfilled). In short, it seems that it does not really make sense to give a satisfaction interpretation of a desire theory unless it is an actual desire theory.

[128]True, it would probably be good for the actual Bert if he would be transformed into this hypothetical person (especially if he has a desire to be rational), but this does not show that it would be good for the actual Bert to do what he would want to do if he were a different person.

[129]Well, not exactly the same: The fact that Bert's brother wants to have friends does not imply that Bert wants his brother to have friends, but that he himself wants to have friends.

[130]Cf. note 129.

brother's, it would not be good for Bert to live the kind of life his brother wants to live. As Griffin (1986) points out,

> [i]t is doubtless true that if I fully appreciated the nature of all possible objects of desire, I should change much of what I wanted. But if I do not go through that daunting improvement, yet the objects of my potentially perfected desires are given to me, I might well not be glad to have them; the education, after all, may be necessary for my getting anything out of them. This is true, for instance, of acquired tastes /.../. *Utility must, it seems, be tied at least to desires that are actual when satisfied* (p 11, my italics)[131].

As I see it, this means that the idealized desire theorist has to give up his original position, but it doesn't necessarily force him to accept the actual desire theory. He can also try to adopt some kind of middle position, viz. by weakening his original claim. There are at least two ways in which this can be done:

(i) The idealized desire theorist might restrict his attention to those cases where the actual person and the hypothetical (rational) person are similar enough to count as "the same person" (in the qualitative sense), e.g., cases where the actual person is a little uninformed or slightly confused (in certain respects) rather than thoroughly irrational or extremely neurotic. He may then claim (more modestly) that it is, *in cases like this*, good for a person (as he is now) to do (etc.) what he would want to do (etc.) if he were fully informed and thinking clearly. This idea is far more plausible than the original idea, but it is still unacceptable. First, it is too vague: how should we determine whether a difference between the actual person and the hypothetical person is "small enough"? Second, it doesn't give us the slightest clue of how we should determine what is good for an irrational person. And third, it is likely that it has (so to speak) inherited its plausibility from the actual desire theory, so why don't we just go all the way and accept the latter theory?

(ii) The idealized desire theorist can also weaken his original position by adopting the following view: "What we should take into account is not /.../ [a] person's ideal preferences for an ideal situation, where he or she has them, but for the actual situation, where he or she is not the

---

[131]Note that this passage has been quoted once before, viz. in note 73 on p 242.

very refined person that he or she [ideally?] is" (Tännsjö (forthcoming), p 88). Is this a plausible view? I think not. Consider the following example[132]: Suppose that the actual Bert has an actual preference for pop music, but that he would prefer classical music if he had his preferences refined. Suppose also that his "ideal preference for the actual situation" is that he listens to "more complex /.../ pop music, which could serve the purpose of educating his /.../ taste [i.e. to make him a more refined person]" (ibid., p 88). That is, Bert wants to listen to simple pop music, his ideal twin brother wants to listen to classical music, and the twin brother (who is deeply concerned about Bert's well-being) wants Bert to listen to complex pop music. Is it, in this case, better for (the actual) Bert to listen to complex pop music than to listen to more simple pop music? Well, it is certainly not nonderivatively better for him, but it might be better for him for instrumental reasons, viz. because it would make him a more refined person. But why should we assume that it is good for the actual Bert to become more refined? On my view, it can only be good for Bert to become more refined if he has an actual higher-order desire to become more refined, i.e. the fact that Bert's brother wants him to become more refined does not give us any reason to assume that it is good for Bert to become more refined.

To conclude, the original version of the idealized desire theory is implausible, and so are the two weaker versions of the theory. This means that argument (2:a) - which is based on the assumption that some idealized desire theory is true - is a bad argument, i.e. it does not constitute any threat to the actual desire theory.

Is there any truth in "the objective list theory"?

The second argument against (D1) (as such) is based on the assumption that there are "certain things [which] are [nonderivatively] good or bad for us, whether or not we want to have the good things, or to avoid the bad things" (Parfit (1984), p 493). So, is this a plausible assumption, or alternatively, is the so-called "objective list theory" (the objectivist form of "non-internalist pluralism") a plausible theory of prudential value? Well, it certainly seems implausible to assume that something to which I have a strong rational aversion can be good for me, but we

[132]The example is originally from Tännsjö (forthcoming), pp 87-88, but has been slightly modified by me.

should not (at this point) rule out the possibility that there are good arguments for the theory. (We will return to this issue in chapter 7).

Scanlon's argument

The next argument purports to show that the desire theory should be totally rejected. The argument given by Scanlon (1993) is a good example of this type of argument[133], so I quote him at full length (note that he seems to take the object interpretation of the desire theory for granted):

> /.../ I now believe that desire theories should also be rejected as accounts of well-being appropriate to the first-person view. I will argue against such theories in the following way. The fact that an outcome would improve a person's well-being ('make his or her life go better') provides that person with a reason (other things being equal) for wanting that outcome to occur. If a desire theory were correct as an account of well-being, then, the fact that a certain outcome would fulfil a person's desire would be a basic reason for that person to want that thing to come about. But desires do not provide basic reasons of this sort, at least not in non-trivial cases. The fact that we prefer a certain outcome can provide us with a serious reason for bringing it about 'for our own sake'. But when it does, this reason is either a reason of the sort described by a mental state view such as hedonism or a reason based on some other notion of substantive good rather than a reason grounded simply in the fact of desire, in the way that desire theories would require. To see this we need to consider each of these cases in a little more detail.
>
> [Part One] In many cases, the fact that I desire a certain outcome provides me with a reason for trying to bring it about because the presence of that desire indicates that the outcome will be pleasant or enjoyable for me. I can have reasons of this kind, for example, for ordering fish rather than tortellini, for climbing to the top of the hill, or for wearing a particular necktie. The end sought in these cases is the experience or mental state which the object or activity in question is expected to produce, and the desire is an indication that this state is likely to be forthcoming (as well as, perhaps, a factor in producing

[133]Another example is the argument given by Tännsjö (forthcoming), on p 89 ff.

it).

[Part Two] In other cases, my desire that a certain state of affairs should obtain reflects my judgement that that state of affairs is desirable for some reason other than the mere fact that I prefer it; it may reflect, for example, my judgement that [that] state of affairs is morally good, or that it is in my overall interest, or that it is a good thing of its kind. This represents, I believe, the most common kind of case in which preferences are cited as reasons for action; the fact that I prefer a certain outcome *is* a reason for action in such a case, but not a fundamental one. My preferences are not the source of reasons but reflect conclusions based on reasons of other kinds. There are, of course, other cases in which I might say that the only reason I have for doing or choosing something is simply that 'I prefer it'. But these cases are trivial ones rather than examples of the typical form of rational decision-making.

My conclusion, then, is that when statements of preference or desire represent serious reasons for action they can be understood in one of the two ways just described: either as stating reasons which are at base hedonistic or as stating judgements of desirability reached on other grounds. What convinces me of this conclusion is chiefly the fact that I am unable to think of any clear cases in which preferences provide non-trivial reasons for action which are not of these two kinds (pp 191-192).

This is not a very good argument, however, and there are several reasons for this:

(1) The main reason why we should not accept Scanlon's argument is that it takes "the immanent perspective on values" for granted. To see what this means, and why it can (and should) not be done, let us first reconstruct the argument as follows:

(P1) The desire theory implies that "the fact that a certain outcome would fulfil a person's desire would be a basic reason for that person [i.e. from his or her own perspective] to want that thing to come about", or to bring it about.

(P2) This implication is (in all non-trivial cases) false. *From the point of view of the desiring subject*, "serious reasons for action" are almost never "grounded simply in the fact of desire", i.e. it is almost never

the mere fact that he desires something that gives him a reason to bring it about.

(C) The desire theory is false.

The reason why the argument does not work is that the first premise is false, which means that the second premise (which is true) does not really give any support to the conclusion.

The reason why (P1) is false is not that it is formulated in terms of reasons for action, but that it makes the desire theory inconsistent with how things appear to us[134]. That is, the premise could also be formulated in evaluative terms, viz. as follows: "The desire theory implies that the mere fact that a certain person desires a certain outcome gives this person a good reason to regard this outcome as good for him". A desire theorist need not adopt the immanent perspective on value, however; he might also accept the idea that "[t]he access to objective value is not given in the immanent perspective but in the transcendental one - in the perspective of the detached spectator" (Rabinowicz and Österberg (1996), p 12)[135].

To get a better grasp on "the immanent mistake", it might be helpful to look at Sen's (1985) "version" of the same mistake. This is how Sen tries to refute the idea that "desire is prior to (prudential) value", i.e. the idea that if a certain situation is nonderivatively good for a certain person, then it is good *because* the person desires that the situation holds[136]: On Sen's view, the role of desire is "evidential" rather than

---

[134]The idea that the desire theory has direct implications for what we have reason to do is also rather problematic, however. It may be true in some "objective" sense that "[t]he fact that an outcome would improve a person's well-being /.../ provides that person with a reason /.../ for wanting that outcome to occur", but it is hardly true if we adopt the immanent perspective. In this case, the person does not have a reason for wanting X to occur unless he (i) believes that X is good for him, and (ii) wants to improve his own well-being.

[135]It is worth noting that the reflective, transcendent perspective is as consistent with "the first-person view" as it is with "the third-person view". If we happen to chose the self-interest theory as our normative context, this does not in any way force us to view things from the immanent perspective. The idea that a certain person has a ("transcendent") reason to fulfil his desire that X is not inconsistent with the idea that a person's ("immanent") reason for promoting X is not the fact that he has the desire but the respect in which X appears to him as desirable (cf. note 97 on p 257).

[136]This idea does not just presuppose that the object interpretation of the desire theory is correct; it is also an essential component of the-theory-under-this-interpretation.

"value-giving". The "activity of desiring" does not *make* the objects of desire have value; instead, desire "gives evidence - *possible* evidence - of value" (p 189). To support this view, he invites us to compare the following two statements:

(I) I desire x because x is valuable for me.

(II) x is valuable for me because I desire x[137].

He then continues:

> The former statement is intelligible and cogent in a way the latter clearly is not. Valuing something is a good reason for desiring it, but desiring something is not an obvious reason for valuing it. /.../ [I]t is hard to escape the conclusion that the main connection between desire and value is the evidential connection in (I) (ibid., p 190).

However, this argument does not really support the idea that the activity of desiring does not have a "value-giving" role. To see why (II) might not be so implausible after all, let us first formulate it in the third person, i.e. as "X is valuable for P because P desires that X". Once we have done this, it is pretty clear that it can be interpreted in two different ways, viz.

(II:a) X is valuable for P (from a transcendent perspective) because P desires that X, and

(II:b) X is valuable for P according to P (from P's perspective) because P desires that X.

(II:b) is, in most cases, a false claim: The mere fact that I intrinsically desire something does (from my perspective) not give me a good reason for regarding it as valuable for me (not even if the desire is relevant)[138]. The following picture is much more plausible: If I have an intrinsic relevant desire that something is the case, then this something

---

[137]Where the formula "A because B" should (of course) be understood as "B is a good reason for A", and not as "B is the cause of A".

[138]But it is not always false, i.e. there are cases where the fact that I intrinsically desire that X gives *me* a conclusive reason to regard X as intrinsically good for me. These are, I think, the cases where my desires cannot really be criticized, i.e. what we normally call "matters of taste" (this is probably the "trivial cases" which Scanlon has in mind). Possible examples of such "value-giving" desires are my desire to have a certain experience when I have it, the love of cold showers, or the aversion to hearing the sound of squeaking chalk (cf. Parfit (1984), p 123).

will (most likely) appear to me as "good" (e.g., as "good for me"), and will do so in virtue of some of its (intrinsic) features. From my point of view, it is the presence of these (apparent) features which explains and justifies my desire. This suggests that what (II:a) really amounts to is that a situation is (or may be) good for a person (from a more objective point of view) because it is good for him from his own (subjective) point of view, and this claim is far from implausible. In any case, neither Sen nor Scanlon gives us any reason for rejecting it: (II:a) does not take P's own perspective for granted, and the fact that (II:b) is often false is (therefore) no reason for regarding (II:a) as false. In short, the argument does not hit the desire theorist who rejects P's own idea of what is going on[139] [140].

(2) Another major weakness in Scanlon's reasoning is that it seems to presuppose atomism. The positive claim he purports to establish is (roughly) that the most plausible theory of prudential value is an "objective list theory" (cf. part two) with a hedonistic component (cf. part one)[141]. But how does he argue for this? Well, the intuitions to

---

[139]At this point, the following comparison with pleasure may be fruitful: The hedonistic counterpart to (II:a) is (i) "X is (instrumentally) good for P because P takes pleasure in X", and the counterpart to (II:b) is (ii) "X is good for P, according to P, because P takes pleasure in X". Now, (ii) is often false, since we often take pleasure in something because we ourselves regard it as intrinsically valuable. But this does not make (i) - the hedonist's claim - false, and the reason for this is (of course) that the hedonist does not adopt P's own perspective.

[140]It is worth noting that Sen's argument suffers from several other defects as well, e.g., the following ones: (a) There is no reason why we should assume that there is any evidential connection between P's desires and what has value-for-P. (I) is not plausible unless it is interpreted as "I desire X because X *appears to me* as valuable for me", but in this case, desire is hardly "possible" evidence of value-for-me; and if we consider the fact that there are many other possible reasons for desiring something, desire is not even evidence of value-for-me-according-to-me. (b) When Sen claims that the fact that P desires that X is not a good reason for regarding X as good for P, it is obvious that he has the (object interpretation of) the *unrestricted* actual desire theory in mind (and so has Scanlon). But what if we exclude all uninformed desires from the class of relevant desire? Is it really implausible to claim that the fact that P has a *rational* intrinsic desire that X is a good reason for regarding X as nonderivatively good for P? I think not.

[141]However, he also he admits that "statements of preference or desire" can represent (that preferences can provide) "trivial reasons for action", and this suggests that he thinks that there is some truth in the desire theory. But it is not clear if this means that it can have nonderivative value for a person to have a certain desire fulfilled, even if this does not make him any happier, and even if the object of the desire is not (in some "objective" sense) nonderivatively good for him.

which he appeals are both about wholes, wholes which include a desire element. If we translate his "reason-talk" into "value-talk", these intuitions can be formulated as follows: In part one, it is suggested that if I desire X, if X occurs (if my desire is fulfilled), and if I take pleasure in X, then it is good for me that X occurs, and in part two, it is suggested that if I desire X, if X occurs (the desire is fulfilled), and if X is good for me in some objective sense, then it is good for me that X occurs. But the fact that these intuitions are both plausible does not support Scanlon's positive claim unless we (implausibly) assume that atomism is true. To see that this is so, let us take a closer look at the intuition in part one. Suppose that it is true that if I desire that X, and if the occurrence of X makes me happier, then it is good for me that X occurs. Suppose also (for the sake of the argument) that if X's occurrence would not have made me happier, then it would not have been good for me. Now, the point is that even if we make all these assumptions, we can not conclude that the *only* good thing about X is that it makes me happier; it is only possible to conclude such a thing if one assumes that "the value of the whole is just the sum of the value of its parts". But there is no reason why we should make this atomistic assumption. There are alternative views which seem more plausible, e.g., the following one: "A situation cannot be valuable for a person unless its existence makes him happier. However, it is nonderivatively better for a person if his happiness is based on real desire-fulfilment than if it is based on a false belief that he had a desire fulfilled. That is, if two wholes contain the same 'amount' of pleasure, then the whole which has real desire-fulfilment as a part is finally better than the whole which has a false belief that a desire is fulfilled as a part. Moreover, it is better for a person to take pleasure in something which he desires than to take pleasure in something to which he has an aversion". (We can apply the same kind of reasoning to show that Scanlon's reasoning in part two is defective).

To conclude: Scanlon has not managed to show that the desire theory is a bad theory, and neither has he managed to show that the "objectivist" alternative which he has in mind is a good theory.

And to summarize section 5.3 as a whole: With the possible exception of (2:b) on pp 274-275, it seems that there are no strong arguments against the rationality-oriented (and awareness-oriented) Success Theory that was presented on pp 271-272 above. There are some good arguments for the theory, however, and this suggests that we have

reason to accept it, but with the following proviso: It remains to be seen whether there are any objective prudential values, i.e. things which are nonderivatively good and bad for us, regardless of what we want and do not want[142] (we will soon turn to this issue).

[142]At this point, it is worth noting that if there are such things, then this suggests that the connection between desire and value is *sometimes* "evidential" (as Sen claims). That is, the fact that someone intrinsically desires a certain thing *may* (in this case) be an indication of the (objective) fact that this thing is good for him.

Chapter Six

"The objective list theory" I

A list of possible prudential values

It is now time to look at the third type of substantive good theory, viz. "non-internalist pluralism" (i.e. Parfit's "objective list theory", conceived of as a type of substantive evaluative theory). As we have already seen (in section 1.2), theories of this type make the following central claims:

(1) There are several (universal) prudential values, i.e. the facts that have nonderivative value for us are of several types.

(2) It is not the case that all the facts that have nonderivative value for a person are internal to this person[1]. (This implies that at least some of the facts that have nonderivative value for P-at-*t* are *not* of the type "P feels pleasure at *t*").

(3) At least some of the "non-internal" facts that have nonderivative value for P-at-*t* are *not* of the type "P has a desire fulfilled at *t*".

(2) and (3) tell us that there are other facts besides pleasure-facts and desire-fulfilment-facts that are nonderivatively good for us, and if we assume that it is the facts themselves that are good for us (i.e. that the value of such a fact is independent of how it is related to other facts)[2], this seems to allow for the possibility that there are facts that are good for us, whether or not we want these facts to obtain, and independently of what we feel about them (and this is, roughly, what "the objectivist" claims).

The fact that none of the claims above are substantive evaluative

---

[1]However, it is rather obvious that every plausible "objective list theory" will attribute value-for-P to certain internal facts about P. See below.

[2]This is what the *pure* versions of non-internalist pluralism assume. But as we will see in chapter 7, on pp 362-366, there are number of "weaker" interpretations of the theory, interpretations which can be invoked if all the strong (pure, "raw") versions are refuted.

claims implies that "non-internalist pluralism" is not a substantive evaluative theory, but a *type* of substantive theory. In fact, the different versions of the "theory" need not ("substantively speaking") have anything in common. Or alternatively put, although every concrete version of the "theory" makes specific claims about what is nonderivatively good and bad for us, no version of the "theory" makes any specific substantive claims *qua* non-internalist pluralist theory. This means that no "objective list theory" can really be assessed *as such*. It is concrete versions of the theory (or better: specific substantive claims about what has prudential value), and nothing else, that can be assessed[3].

So, in order to find out whether the most plausible theory of prudential value is of the non-internalist pluralist type (i.e. whether there are any prudential values that fit the description in (1)-(3) above), what we need to consider are "concrete" versions of the "theory". Or better: We first need to look at what types of relational or external facts non-internalist pluralists have actually regarded as prudentially valuable, i.e. what items they have actually included in their positive and negative "lists"[4]. We can then ask, for each suggested type of fact, whether it is really plausible to attribute prudential goodness (or badness) to facts of this type. This is the type of investigation that can give us an idea of what (if any) relational or external facts that have nonderivative value for us, i.e. of what the most plausible "list" is like[5].

In this chapter, we will look at some of the substantive evaluative claims which have actually been made by various pluralists. The purpose of the chapter is merely to generate a list of possible prudential values, i.e. a list of candidates to the title "non-internal facts which have non-

---

[3]In this respect, it differs from hedonism and the satisfaction interpretation of the actual desire theory. There are (as we have seen) several versions of two latter theories as well, but all these versions share a common substantive content, and it is therefore much more appropriate to regard these two theories as real substantive theories.

[4]It would also be of interest to see how the pluralist would answer the following questions, viz. (i) "Can the situation-types to which he attributes prudential value be ranked in any way, and if they can, how should they be ranked?" and (ii) "What type of combination of prudential values is best for a person?". I haven't come across any pluralist who has really tried to answer these questions, however.

[5]This means that there is really no need to treat (i) "What possible version of the theory is most plausible?" and (ii) "Is this theory a plausible theory of prudential value?" separately.

derivative value for human beings". In chapter 7, we will turn to the question of plausibility, viz. we will ask whether any of the substantive claims which non-internalist pluralists have made are (in fact) plausible. A central question here is (of course) what kinds of arguments that can be given for such claims, e.g., for the idea that it is good for all of us to have friends. Can any universal substantive claims of the form "All non-intrinsic facts of type X[6] are nonderivatively good (or bad) for all human beings (at all times)" be justified, and if so, how? In particular, what would an acceptable *subject-oriented* justification of such a claim look like, e.g., what is it about us (about our "constitution") that makes it nonderivatively good for all of us to have friends, or to be engaged in creative activity[7]?

A list of possible prudential values

The list of possible human goods includes such different things as "moral goodness, rational activity, the development of one's abilities, having children and being a good parent, /.../ the awareness of true beauty" (Parfit (1984), p 499), "health, mental and physical functioning, enjoyment, personal achievement, knowledge or understanding, close personal relationships, personal liberty or autonomy, a sense of self-worth, meaningful work, and leisure or play" (Sumner (1996), p 180)[8]. And prudential badness might be attributed to such things as "being betrayed, manipulated, slandered, deceived, being deprived of liberty or dignity, and enjoying either sadistic pleasure, or aesthetic pleasure in what is in fact ugly" (Parfit (1984), p 499), or to immoral behaviour, passivity, loneliness, or being ashamed to appear in public[9].

---

[6]Where X is not desire-fulfilment (or aversion-fulfilment).

[7]Here, we should (unlike Sumner (1996)) allow for the possibility that a unified account of what makes things good and bad for us may not be possible, and that the relevant "justifications" may (therefore) vary from case to case.

[8]It is important to notice that many of the things on this "positive list" (e.g., enjoyment, play, and meaningful work) contain pro-attitudes as constituent parts, and that it can (for this reason) be rather trivial to claim that such things are good for us. But even these claims have some content; it is, after all, possible to reject them.

[9]At this point, it is worth noting that every list of prudential values is (at least to a certain extent) relative to some world view, and that the world view that seems to be presupposed here is (for the most part) "modern, Western, and atheist" (cf. Griffin (1996), p 150, note 19). For example, none of the values listed can be regarded as "spiritual".

Now, there is really no point in striving for completeness here, so I what I will do is to restrict my attention to those positive items[10] that seems (to me) most important. These "intuitively important" items will be classified into seven groups, viz. (1) activities and other "agent-goods", (2) social and relational goods, (3) experiences and other mental states[11], (4) to be (*qua* experiencing and thinking subject, or *qua* "patient") in contact with reality, (5) to be a certain kind of person and to live one's life in a certain way (to function in a certain way), (6) personal development, and (7) freedom and other "potentialities"[12].

(1) Activities and other "agent-goods"

There are at least four different kinds of claims which can be subsumed under this heading, viz.:

(a) The claim that it has value for us to be active at all, or better, that it is (on the whole) nonderivatively better for us to be active than to be passive.

(b) The claim that it is good for us to be engaged in certain kinds of activities, or to perform certain kinds of actions, rather than others. For example, it has been claimed that it is good for us to act morally; to do constructive or productive things; to be engaged in play, or meaningful work, or contemplation, or political activity; or to exercise our respective abilities to a high degree.

(c) The claim that it is important for us *how* we do what we do, i.e. that it is good for us to do what we are doing "in a certain way", e.g.,

---

[10]That is, I will (like almost everybody else in this field) have little or nothing to say about what situations that are nonderivatively bad for us (and why). But as a rule of thumb, we can always *assume* that the opposites of the good things are bad for us, and once we have seen the reasons that can be given for regarding the good things as good (assuming that there are such reasons), it will probably be easy to see what it is that makes the bad things bad.

[11]The facts which belong to this category are (of course) internal rather than relational or external, but they deserve to be included anyway: A pluralist theory which excludes such facts altogether from its list is, after all, hardly a plausible theory of prudential value (cf. note 1 above).

[12]This classification is (of course) not the only possible classification; it is not just that different categories may be used, there are also alternative ways in which the items can be classified (given a certain set of categories). Moreover, the present classification does not purport to be complete, and all the categories used are probably not mutually exclusive. What matters is that the classification is good enough for my modest purpose, which is to make things a little more systematic than they would otherwise have been.

in a certain spirit, or with a certain quality, style or attitude. It has, for example, been claimed that it is good for us to do things in a rational manner, or to be creative, or to perform with "perfection" (i.e. to do what one is doing well, or skilfully)[13].

(d) The claim that it is good for us that our actions have certain results, or that we "get certain things done". It has, for example, been claimed that it is important for us to "make a difference", i.e. that the things we do have an effect, an impact beyond themselves (which is preferably valuable); or that it is good for a person to "do something with his life", or to "accomplish things with his life", in the sense of "achieving something valuable" (cf. Griffin (1996), who is somewhat more specific when he focuses on "the kind of achievement that would save a life from futility", "the achievement of the kind of value that gives life a weight or point" (p 54)).

(2) Social and relational goods

Following Aristotle, we can divide the relational goods into two types, viz. (a) intimate personal relationships, and (b) social and political relations[14].

An example of a claim which can be subsumed under the second heading is the Aristotelian idea that it is good for a person to belong to a community (to stand in a relation of "civic friendship" to other citizens), and to participate actively in the political life of this community[15].

---

[13]In my view, some ways of doing are best viewed as aspects of "(total) ways of functioning", where a person's way of (mental) functioning also includes how he perceives (e.g., what he notices and does not notice), how he reacts emotionally, how he "copes" (what "coping strategies" he uses), how he thinks, how he relates to other people, how he makes decisions, and so on. If a certain way of doing is best conceived of as one of several aspects of a certain (total) way of functioning (e.g., as in the case of autonomous acting, or authentic acting), I will not treat it here, but in (5) below.

[14]According to Nussbaum (1986), Aristotle claims that there are two kinds of relations that are "essential and valuable parts of the good human life", viz. "social/political relations and *philia*" (p 344), and according to Cooper (1980), "Aristotle holds not only that active friendships of a close and intimate kind are a necessary constituent of the flourishing human life but also that 'civic friendship' itself is an essential human good" (p 303).

[15]Or as Nussbaum (1986) puts it, "Aristotle believes the political participation of the citizen to be itself an intrinsic good or end, without which a human life, though flourishing with respect to other excellences, will be incomplete" (ibid., p 349). However, it is (as I see it) not easy to tell what should count as "political

If we turn to the claim that it is good for us to have intimate personal relationships to other people, we can see that it comes in several different versions, e.g., it might be claimed that it is good for a (mature, adult) person to have children (and to be a good parent), or to have friends, or a lover, or a partner with whom a romantic "we" is formed, or the like[16]. The most common claim of this type is probably the idea that when personal relations "become deep, authentic, reciprocal relations of friendship and love, then they have a value apart from the pleasure and benefit they give" (cf. Griffin (1986), pp 67-68). To get a deeper understanding of what may be involved in a claim of this type, we will take a closer look at one of the most important Aristotelian prudential values, namely *filia* (or *philia*)[17].

The special case of filia; its nature and its forms

So, what is *filia* (or "friendship", as it is sometimes translated)? "To be friends, /.../ [two persons] must be mutually recognized as bearing goodwill and wishing well to each other", says Aristotle (in *NE*, VIII.2, on p 194). Or as Cooper (1980) puts it: "According to /.../ [Aristotle], *filia*, taken most generally, is any relationship characterized by mutual liking /.../, that is, by mutual well-wishing and well-doing out of concern for one another" (p 302). That is, "without mutuality of genuine well-wishing for the other person's own sake the relationship will not deserve the title of *philia* at all" (Nussbaum (1986), p 355). Nussbaum also adds independence to the list of requirements for *filia*; "the object of *philia* must be seen as a being with a separate good, not as simply a possession or extension of the *philos*; and the real *philos* will wish the other well for the sake of that separate good" (ibid., p 355).

---

participation" in a modern society like ours.

[16]It is also possible to attribute prudential value to certain parts (or aspects) of these prudentially valuable relationships, e.g., to some of the attitudes that are involved. It can, for example, be claimed that it is good for us to be liked, loved, needed, respected, recognized, admired, taken into account, or regarded as important by (significant) others, or that it is good for us to love or recognize others (even when these attitudes are not mutual). It is possible that honour (a central Aristotelian value) is of this type, viz. an attitudinal component in (good) "civic friendship".

[17]If we want to be more precise and specific, the Aristotelian view on the matter can be formulated as follows: (i) It is nonderivatively good for us to have friends (to have friends is nonderivatively better than to be alone), and (ii) character-friendship is the most valuable kind of friendship (it is probably the only form of *filia* that can really be regarded as a *necessary* constituent of a good human life).

This is how she sums up Aristotle's view on *filia*:

> *Philia* requires, then, mutuality in affection; it requires separateness; it requires mutual well-wishing for the other's own sake and /.../ mutual benefiting in action, insofar as this is possible /.../. Aristotle completes his general sketch of *philia* by adding that there must be mutual awareness of these good feelings and good wishes /.../ (ibid., p 355)[18].

Depending on what it is that attracts and binds the one friend to the other (or what type of "reasons" the friends have for wishing each other well), there are (on Aristotle's view) three basic kinds of "friendship", viz. advantage-friendship, pleasure-friendship, and virtue-friendship (or character-friendship).

Advantage-friendship is the type of friendship which is based on some advantage, that the one gets from the other, and advantage-friends are characterized as friends "who love each other for their utility" or "in virtue of some good they get from each other" (*NE*, VIII.3, p 195).

Pleasure-friendships are (instead) "cemented by" the pleasure that the one friend gets from the other; pleasure-friends are friends "who love for the sake of pleasure", i.e. "for the sake of what is pleasant *to themselves*" (ibid., VIII.3, p 195).

That is, in both advantage-friendship and pleasure-friendship, "it is not as being the man he is that the loved person is loved, but as providing some good or pleasure" (ibid., VIII.3, p 195). But it should be noted that friendships of these kinds are not "wholly self-centered: /.../ [they] are instead a complex and subtle mixture of self-seeking and unself-interested well-wishing and well-doing" (Cooper (1980), p 305).

The third type of *filia* is also (according to Aristotle) the best, and it therefore deserves to be characterized in more detail. This is how Aristotle himself characterizes the *virtue-friendship*, or *character-friendship* (as Cooper prefers to call it):

> Perfect friendship is the friendship of men who are good, and alike in virtue; for these wish well alike to each other *qua* good, and they are

---

[18]All these "mutualities" do not imply equality, however. In fact, *filia*-relationships need not be equal, even though they often involve equality (cf. *NE*, VIII.6, p 202). This is but one of several reasons why we should regard the Aristotelian notion of *filia* as much wider than our notion of friendship.

> good in themselves. Now those who wish well to their friends for their sake are most truly friends; for they do this by reason of their own nature and not incidentally; therefore their friendship lasts as long as they are good - and goodness is an enduring thing (*NE*, VIII.3, p 196).

That is, this type of friendship is based on the mutual recognition of the other person's goodness (excellent character), and this seems to imply that "for their own sake clearly only good men can be friends; for bad men do not delight in each other unless some advantage come of the relation" (ibid., VIII.4, p 198). According to Cooper (1980), this form of *filia* is not "the exclusive preserve of moral heroes", however:

> Some virtue-friendships might involve the recognition of complete and perfect virtue /.../; other friendships of the same type might be based not on the recognition by each of perfect virtue in the other but just on the recognition of some morally good qualities that he possesses (or is thought to possess) (p 306).

For this reason, Cooper thinks that "friendship of character" is a more appropriate name for this kind of friendship than "friendship of virtue (of the good)": "The expression 'character-friendship' brings out accurately that the basis for the relationship is the recognition of good qualities of character, without in any way implying that the parties are moral heroes" (ibid., p 308). This is how Cooper conceives of the relation between two character-friends:

> Each, loving the other for his good qualities of character, wishes for him whatever is good, for his own sake, precisely in recognition of his goodness of character, and it is mutually known to them that well-wishing of this kind is reciprocated /.../. They enjoy one another's company and are benefited by it /.../ and in consequence spend their time together or even live with one another. /.../ [S]uch a friendship, once formed, will tend to be continuous and permanent, since it is grounded in knowledge of and love for one another's good qualities of character, and such traits, once formed, tend to be permanent (ibid., pp 308-309).
>
> A character-friend loves his friend because of properties which belong to the friend essentially and not merely incidentally. This means that he loves him for what he himself is and not for merely

external properties or for relations in which he stands to other persons (ibid., p 315).

And this is how Nussbaum (1986) characterizes the same type of *filia*:

> The central and best case of love between persons is that of love based upon character and conception of the good. Here each partner loves the other for what that other most deeply is in him or herself /.../, for those dispositions and those patterns of thought and feeling that are so intrinsic to his being himself that a change in them would raise questions of identity and persistence (p 356).

She also adds another requirement for "the best type of love", namely that the two *filoi* "must 'live together', sharing activities both intellectual and social, sharing the enjoyment, and the mutual recognition of enjoyment, that comes of spending time with someone whom they find both wonderful and delightful (ibid., pp 357-358)[19]. She also points out that "Aristotle means by 'living together' something more than regular social visiting: if not residence in the same household, then at least a regular, even daily association in work and conversation (ibid., p 358).

This ends the section on *filia*.

(3) Experiences and other mental states

A non-internalist pluralist may of course attribute prudential value to things like "pleasures, the perception of beauty, absorption in and appreciation of nature - the enjoyment of the day-to-day textures of life" (cf. Griffin (1986), p 67). All the claims which can be subsumed under this heading do not have this hedonistic flavour, however, e.g., there is also the idea that it is nonderivatively good for a person to have a sense of self-worth, or to "feel" self-respect.

(4) To be (qua experiencing and thinking subject) in contact with reality

The general claim here is that it is nonderivatively good for a person to be in contact with (internal and external[20]) reality, or to have "authentic

---

[19]This means that it was character-*filia* that Aristotle had in mind when he wrote that "there is nothing so characteristic of friends as living together" (*NE*, VIII.5, p 200), and that spending one's days together and delighting in each other "are thought the greatest marks of friendship" (*NE*, VIII.6, p 201).

[20]The reality that Nozick (1989) has in mind is primarily external reality. He

experiences". There are, however, many "versions" of this claim, versions which correspond to the many ways in which one may be in (or out of) touch with reality. Examples of such (more specific) claims are:

(a) "Simply knowing about oneself and one's world is part of a good life. We value, not as an instrument but for itself, being in touch with reality, being free from muddle, ignorance, and mistake" (Griffin (1986), p 67). That is, it is good for a person to have true beliefs about himself and his world (rather than to have false beliefs, or to be ignorant), to understand how things work, and how they are connected to each other.

(b) It is good for a person to perceive things "as they are", free from illusion and delusion, e.g., to be in an enlightened state.

(c) A special case of (b) which is often mentioned is this: It is good for a person to be aware of true beauty (assuming that there is such a thing), i.e. (roughly) to take aesthetic pleasure in what is "in fact beautiful" (rather than in what is "in fact ugly").

(d) It is good for a person that his emotions (especially his so-called "import-attributing emotions") are appropriate. (As I see it, the idea that it is good for a person to have a "sense of proportion" can be regarded as an important part (or aspect) of this claim)[21].

(5) The prudential value of being a certain kind of person and/or living one's life in a certain way

There are two general claims which can be subsumed under this heading, viz. (a) the idea that it has nonderivative value for a person to be a certain kind of person, or alternatively, that a person's well-being is directly dependent on what kind of person he is, and (b) the idea that it is nonderivatively good for a person to live (or "lead") his life in a certain way, or to function in a certain manner[22].

---

writes: "To focus on external reality, with your beliefs, evaluations, and emotions, is valuable in itself, not just as a means to more pleasure or happiness" (p 106).

[21]It is worth noting that some of these ways of being in touch with reality, especially (b) and (d), can be regarded as "defining characteristics" of the well-functioning person, and that they could (for this reason) also have been included in (5).

[22]An example of a philosopher who believes that a person's level of well-being depends both on what kind of person he is and how (in what way) he lives his life is Nozick (1981). In his view, the value of a person's "existence" (i.e. a person's well-being) is not just a function of what kind of life he has, but also of what kind of person he is, and how these two are connected to each other, or alternatively,

(a) There are two major versions of the first claim, viz. (i) the idea that it is nonderivatively good for a person to have a certain kind of body, or to possess certain bodily features (e.g., to be healthy, beautiful, or well-built), and (ii) the idea that it is good for a person to have a certain kind of mind, i.e. to possess certain character traits, skills, abilities, talents, attitudes, dispositions, inclinations, sensitivities, habits, motivational features, or the like (e.g., to be courageous, or psychologically well-adjusted, or morally good, or intelligent, or good at what one is doing)[23].

(b) Before we take a closer look at the second type of claim, let us first say something about what it is for a person to live (or lead[24]) his life in a certain way, or to function in a certain way. We can think of a person's living (or functioning) as a "process, which /.../ occurs in, though not necessarily inside, the person, and /.../ [which] issues in his life" (Wollheim (1984), p 2), a process which is determined by two kinds of factors, viz. what kind of person he is like (what kind of mind he has), and what his environment is like. It may also be helpful to think of livings as *ways* of living, or *ways* of functioning, where we can (roughly) conceive of a person's overall way of functioning as "the *spirit* in which he approaches life, meets its trials, and decides on goals and action" (to borrow a phrase from Griffin (1986)), or perhaps as his "life-style", "coping-style", or "posture".

We should also add that the process which is a person's living is best regarded as "multi-dimensional". There are many aspects of a person's living, e.g., our actings, relatings (or interactings), emotional reactings, perceivings, thinkings (reasonings), rememberings, and fantasizings can

---

how he lives (or "leads") his life. Or expressed in superlative terms (Nozick has a tendency to express himself in this way), a person's existence is (on the most general level, that is) the most valuable kind of human existence if and only if (i) he has the best kind of life, (ii) he is the most valuable kind of person (supposing that there are such things as *the* best kind of life and *the* best kind of person), and (iii) there is a proper interconnection between the kind of person he is and the kind of live he lives; the person is "leading the most valuable life, or living it" (p 413), i.e. having the best kind of life stems (in the right way) from being the best kind of person (cf. ibid., p 412).

[23]With the exception of (1:c), this is where we find almost all so-called "perfectionist values" (which can be of relevance in this context, that is).

[24]I prefer the term "living" to the term "leading", however, and the reason for this is that the latter term suggests that we are (to a considerable extent) in control of how our lives are lived, something which is (in many cases) not true.

all be regarded as aspects of our total functionings[25].

The general claim that it has nonderivative value for a person to live his life (function) in a certain way comes in many forms, e.g., it may be claimed that it is good for a person to live autonomously, authentically, rationally, creatively, mindfully, self-expressively, decisively, or the like. In some of these cases, prudential value is attributed to certain ways of living "as wholes" (e.g., to autonomous, rational or authentic living), but in other cases, the focus is on one or a few dimensions only, e.g., as when value is attributed to creative thinking, self-expressive acting, committed relating, or mindful perceiving.

Now, it seems much more plausible to attribute nonderivative value-for-P to the fact that P is living (functioning) in a certain way than to the fact that he is a certain kind of person. In fact, I don't think there is any personal feature (bodily or psychological) that has nonderivative value for all the persons who possess it[26]. The reason why no psychological feature has nonderivative value for its "possessor" is that all features of this type are mere potentialities, and as such, they can only have derivative value[27]. My belief that purely bodily features can not have nonderivative value is based on a different idea, viz. the idea that something can have nonderivative value only if it somehow involves experiences (or experiencing). We might also point out that "living-oriented" conceptions of the human good have always been much more common than "person-oriented" conceptions. For example, Aristotle (cf. note 27), a number of Eastern philosophers (like Taoists and Zen Buddhists), and the Existentialists have all been focusing on the living (on what it is to live one's life in a good way).

To get a fuller understanding of what might be involved in "living-oriented claims", we will now take a closer look at two claims of this type, viz. (1) the idea that it is nonderivatively good for a person to live his life in an autonomous way, and (2) the Aristotelian idea that it has

---

[25]Cf. note 13 on p 290 above.

[26]It is worth noting that this idea is not necessarily inconsistent with the object interpretation of the Success Theory, which implies that if *I* have an intrinsic desire to have a certain personal feature, then it is also nonderivatively good for *me* to have this feature. (It is, after all, rather unlikely that there is a feature F such that everyone who has F also has an intrinsic desire to have F).

[27]Roughly speaking, a potentiality has derivative value if its realization has value; cf. the Aristotelian view that what has nonderivative value is not possessing the virtues, but exercising them.

nonderivative value for a person to function in accordance with excellence (virtue, or reason).

(5:1) Autonomy

Here, the claim is that it is good for a person that his living is (in some strong sense) "his own", where this claim is normally specified as follows: It is good for a person to choose his "own course through life, making something of it according to /.../ [his] own lights" (Griffin (1996), p 29)[28], or alternatively, to be the kind of person who does (with respect to personal matters, that is) what he himself decides to do, and who decides to do what he decides to do because he wills it, i.e. because he actually wants to do it (cf. Tännsjö (forthcoming), pp 97-98). The reason why the claim is specified in this way is that autonomous living is regarded primarily as a way of acting (or doing) and a way of deciding on goals and action. There may well be more to full autonomous functioning than autonomous acting and deciding, however, e.g., it may also involve autonomous desiring and believing (cf. Elster's genetic theory, e.g., on pp 253-255), autonomous thinking, and the like[29].

(5:2) Functioning in accordance with excellence

The Aristotelian idea that it has nonderivative value for a person to function in an excellent way, in accordance with "complete virtue", can be divided into two parts: (i) It is good for a person to have an excellent character (to possess *ethike arete* in the singular) *and* to function accordingly. This claim can also be (roughly) formulated as follows: It is good for a person to function in accordance with the excellences of character (in the plural) and *phronesis* (or practical wisdom). (ii) It is good for a person to possess an excellent intellect (in the singular), i.e. to be a rational person, *and* to function accordingly, i.e. rationally. Or more specifically, it is good for a person to function in accordance with the excellences of intellect (in the plural), especially in accordance with

---

[28]In Griffin's view, this kind of living "is at the heart of what it is to lead a human existence" (ibid., p 29), where the term "human" should be understood in "the special normative sense that /.../ centres on 'agency'" (ibid., pp 29-30).

[29]It may also be claimed that full autonomous functioning is nothing but autonomous acting and deciding, but that an action or decision is not fully autonomous unless it is based on (or caused by) desires and beliefs which are themselves autonomous.

sophia (or philosophic wisdom)[30].

Now, in order to get a fuller understanding of these two claims, we need to know what it is to possess the excellences of character (excellences of intellect), and what it is to exercise (function in accordance with) these "virtues"[31].

To function in accordance with ethike arete

The excellences of character listed by Aristotle - e.g., courage, temperance, and liberality (generosity) - all belong to the category of *hexis*: they are all states of character, or permanent dispositions, rather than, for example, skills or capacities. So, what do the "moral virtues" dispose their possessor *to*? Well, a person who possesses a certain excellence of character (e.g., courage) is of course (by definition) disposed (a) to act in a certain way (e.g., courageously), but also (b) to react (emotionally) in a certain way (e.g., in a courageous manner, viz. to feel fear in an appropriate way, and so on)[32], and (c) to take pleasure in certain things (especially in acting in accordance with the excellence in question) and to be annoyed at other things[33]. (And maybe we can also add that the virtues dispose their possessors (d) to perceive things, or to be atten-

---

[30]That is, it is not sufficient that one possesses the different excellences (or "virtues"); one also has to use (or exercise) them. Aristotle's main reason for placing the chief human good in use rather than in possession is that mere possession of virtue may (*qua* "state of mind") "exist without producing any good result" (*NE*, I.8, p 16), e.g., that it "seems actually compatible with being asleep, or with lifelong inactivity, and further, with the greatest sufferings and misfortunes" (cf. ibid., I.5, p 7).

[31]That is, we will follow Aristotle's advice, when he says that "[s]ince happiness [*eudaimonia*] is an activity of soul in accordance with perfect virtue [*arete*], we must consider the nature of virtue; for perhaps we shall thus see better the nature of happiness [*eudaimonia*]" (*NE*, I.13, p 24).

[32]That is, virtue is not just a *hexis* concerning action (*praxis*), it is also a *hexis* concerning reaction (*pathos*). The *pathos*-element in *ethike arete* (as a whole, in the singular) can be characterized in the following way: A person who possesses an excellent character does not just act in an appropriate way, he also feels and manifests each emotion at a proper time, on proper matters, for the right reason, in the right way, and so on.

[33]That is, there is (on Aristotle's view) an essential connection between virtue, on the one hand, and *hedone* and *lupe* (liking and disliking, pleasure and pain), on the other. That an agent takes pleasure in acting in accordance with a certain excellence (e.g., courage) is a sign that he possesses this excellence, and to enjoy, or like, doing virtuous acts (in general) is a sign that the virtuous disposition (as a whole, in the singular) has truly been acquired (cf. *NE*, I.8, pp 16-17).

tive, in certain ways rather than others)[34].

The fact that the moral virtues are exercised in a number of different "dimensions" means (roughly) that a person is functioning in accordance with *ethike arete* if and only if he is acting, reacting (emoting), liking and disliking, and perceiving (attending) in accordance with all the different excellences of character (in the plural)[35].

To function in accordance with the excellences of intellect

Of the five excellences of intellect listed by Aristotle, two are clearly practical; *phronesis* (which is regarded as a "true and reasoned state of capacity to act with regard to the things that are good and bad for man" (*NE*, VI.5, p 142)), and *techne* (which is regarded as "a state of capacity to make [in contrast to act], involving a true course of reasoning" (ibid., VI.4, p 141)). The major excellence of the theoretical intellect ("the best part of our soul") is *sophia*, or philosophic wisdom. In fact, it seems that *sophia* should be regarded as identical with the complete excellence of the theoretical intellect; Aristotle says explicitly that it is the combination of the remaining two excellences of intellect, i.e. *nous* (or "intuitive reason"), and *episteme* (or "scientific knowledge"). This means that a philosophically wise person is (roughly) a person who possesses the truth about the first principles, and who knows what follows from these first principles.

But what kind of thing is *sophia* (the finest of the excellences of intellect)? Well, it is not a *hexis* (a state of character or permanent disposition), but (rather) a kind of "knowledge". This is supported by

---

[34]There are of course many other things which can be said about the Aristotelian conception of the excellent character, e.g., (i) that every "moral virtue" is connected to a sphere of human experience "where human choice is both non-optional and somewhat problematic" (Nussbaum (1988), p 37), or (ii) that every excellence of character is a *hexis* that is "in the mean", or (iii) that the excellences of character form a unity, i.e. that the possession of one excellence of character presupposes the possession of all the others, or (iv) that overall excellence of character (in the singular), as well as every single excellence of character, presupposes *phronesis* (practical wisdom), and vice versa. None of these features of the Aristotelian conception is of any importance in this context, however.

[35]And since all functioning in accordance with "complete virtue" also involves the exercise of *phronesis*, we might also add that all excellent functioning is (on Aristotle's view) rational functioning, e.g., it is not just good activity but "good activity *according to, shaped by*, the work of reason" (cf. Nussbaum (1986), p 376). That is, we can also say that person is functioning in accordance with *ethike arete* if and only if he is acting, reacting (etc.) in a *rational* way.

some of the things that Aristotle says about it, e.g. that it is "the most finished of the forms of knowledge", and that it has objects (viz. the highest objects). Perhaps it may also be regarded as a capacity, but I'm not really sure about this.

Now that we have some idea of what it is to *possess* the excellences of intellect; what is it to *function* (e.g., to be active) in accordance with these excellences? Well, to function in accordance with the excellences of the "practical intellect" (with *phronesis* and *techne*) is to put these capacities to use, i.e. to act in a rational way (in accordance with the excellences of character) and to make (produce) things in a rational manner. But what is it to be active in accordance with *sophia*, i.e. how is *sophia* "exercised" or "actualized"? It is clear that sophia is typically manifested in the activity of *theoria*, or pure intellectual contemplation of the highest objects (eternal truths), but it is less clear whether there is any other way in which *sophia* can be actualized.

To sum up, it seems that the idea that it has nonderivative value for a person to function in accordance with the Aristotelian excellences ("moral" and intellectual) can be spelled out as follows: It is nonderivatively good for a person (1) to act in accordance with the excellences of character (and *phronesis*), e.g., courageously (for its own sake); (2) to react emotionally in accordance with the excellences of character, e.g. (in the case of courage), to feel and manifest fear in appropriate ways; (3) to feel pleasure and displeasure (like and dislike) in accordance with the excellences of character, e.g. (in the case of generosity), to take pleasure in giving; (4) to perceive (focus, be attentive) in accordance with the excellences of character, e.g. (in the case of generosity), to be attentive to other people's needs; (5) to produce things in accordance with *techne*, i.e. in a rational manner; and (6) to think or reason in accordance with *sophia*, e.g., to engage in *theoria*.

(6) Personal development

It might also be claimed that it is good for a person to "realize himself", or his "potential", e.g., to realize his talents or develop his abilities, to become a morally better person, to become a more well-functioning person (in the Aristotelian sense), or to become more autonomous[36].

---

[36]It is worth noting that this idea can only be regarded as different from (5) if the prudential value of "becoming better" is (in some way, and to some extent) independent of the value of "being good", i.e. if "the process" does not inherit all its

(7) Freedom and other "potentialities"

According to Sen (1992), there are also prudential values in the "realm of the possible". In his view, a person's "achieved well-being" does not just depend on how he actually lives, it is also (directly) dependent on his "capability to function", viz. on how large his "capability set" is:

> Choosing may itself be a valuable part of living, and a life of genuine choice with serious options may be seen to be - for that reason - richer. In this view, at least some types of capabilities contribute *directly* to well-being, making one's life richer with the opportunity of reflective choice (p 41).

If one takes a closer look at what Sen actually says, one can see that what he really regards as prudentially valuable is not certain isolated capabilities (in the plural), but "choosing", "being able to choose", and "freedom". As Sen himself puts it, "[a]cting freely and being able to choose are /.../ directly conducive to well-being, not just because more freedom makes more alternatives available" (ibid., p 51)[37].

This concludes our list of possible prudential values. However, in what follows, we will have to restrict ourselves to a few of these values. The values I have chosen are: Achievement, *filia* and other intimate relationships, contact with reality, autonomy, functioning well (in the Aristotelian sense), and freedom. My reason for including these particular values is either that they are intuitively plausible, or that it has been possible to find interesting arguments for regarding these things as prudentially valuable. So, let us now turn to the questions of plausibility.

---

value from "the product" (or "result").

[37]Cf. Nozick (1989), who suggests that one of bad-making features of experience machine living is that "once on the machine a person would not make any choices, and certainly would not choose anything *freely*. One portion of what we want to be actual is our actually (and freely) choosing, not merely the appearance of that" (p 108).

Chapter Seven
"The objective list theory" II
A critical discussion of the different non-internalist pluralist claims

The main purpose of this chapter is to find out whether any theory of the non-internalist pluralist type is a justified (or well-founded) theory of prudential value, i.e. whether there are good reasons for regarding any such theory as plausible. Or more specifically, we want to find out whether any of the substantive claims which non-internalist pluralists have made are plausible. For example, is it plausible to claim that it has nonderivative value for all of us to achieve things, or to have *filia*-relationships with other people, or to be in contact with reality, or to live our lives autonomously, or to function well (in the Aristotelian sense), or to be free (able to choose)? Is it possible to justify (or refute) universal evaluative claims of this type? If it is, how can they be justified (refuted)?

To be able to answer questions like these, we have to look at a number of arguments that have been (or can be) given for and against such claims, and then ask ourselves whether these arguments are good arguments. We will start with the counter-arguments.

Arguments against the "theory"

There are two types of arguments against non-internalist pluralist theories, viz. general arguments which are directed against all theories of this type, and more specific arguments which are directed at some specific non-internalist claim, where most of the arguments we will look at are of the former type, i.e. they purport to show that *no* claim of the relevant type is valid. It is also worth noting that all the arguments against non-internalist pluralism which we will look at are also arguments for some alternative theory, either for hedonism or for the desire

theory; these arguments are (so to speak) either coming from the direction of hedonism or from the direction of the desire theory (which is not to say that they presuppose the truth of any alternative theory, however). Let us now look at the arguments themselves:

(1) "As we have already seen (on p 286 above), all pure non-internalist pluralist theories seem to allow for the possibility that there are situations which are good for us, regardless of whether we want these situations to obtain, and regardless of how of we feel about them. For example, such a theory might allow for the possibility that it is good for a person to live autonomously, even if he doesn't have the slightest desire to do so, and even if he suffers from it. This is not a plausible view, however, and this has nothing to do with autonomy in particular. It doesn't really matter if we exchange autonomy for some other alleged prudential value (like friendship or achievement); the implausibility of the view is not affected by such a substitution. This strongly suggests that there are no universal prudential values besides pleasure and desire-fulfilment"[1].

So, is this a good argument? Well, the non-internalist pluralist can (of course) object to the argument as follows: "It is true that if a person has no desire whatsoever to live autonomously, and if he (in addition) suffers from it, then it is, *on the whole*, not good for him to live his life in this way. But this does not imply that there is *no respect* in which it is good for him to live autonomously; in fact, it is good for him in one particular respect, viz. with regard to autonomy!". I don't really know how to respond to this reply. Suffice it to say that in my view, (1) is both valid and relevant[2] (but the argument against (1) is not bad either).

(2) The "theory" also seems to allow for the possibility that a person's well-being can be directly affected by things he doesn't know anything about[3]. As an illustration of this, consider Nozick's (1981) claim that it is *ceteris paribus* better for a person to be moral than to be immoral, or alternatively, that "there is a cost to immoral behavior", viz. a value

---

[1]This argument can be regarded as a mixture of pro-argument (2) for hedonism on p 109 and the second pro-argument for the desire theory on p 195.
[2]However, if someone would not agree with me on this point, there is really nothing more I can say.
[3]Cf. the discussion on pp 266-269 above.

cost, or "value sanction" (cf. p 409). In Nozick's view, this "value sanction" does not depend on whether the agent cares about value (and morality) or not:

> The immoral person thinks he is getting away with something, he thinks his immoral behavior costs him nothing. But that is not true; he pays the cost of having a less valuable existence. He pays that penalty, though he doesn't feel it or care about it. Not all penalties are felt.
>
> /.../ [H]is not caring about value is also part of the cost he is paying: not caring about value is itself something that diminishes his value. Even as he skips happily away, he pays the penalty. The immoral person is not getting away with anything; his getaway attempt itself has a value cost. There is a penalty even if he doesn't realize it or care (ibid., pp 409-410).

Now (so the argument goes), to claim that a person's well-being can (in this way) be directly affected by situations which do not in any way "enter" or "affect" his experience, this is not plausible.

Is this a good argument? Well, this depends on what it purports to show. If it is directed against certain specific non-internalist claims, it is not a bad argument, but it is not an argument which hits all claims of this type, and it does not seem to hit the most plausible claims of this type. For example, it is hardly possible to be totally unaware of the fact that one is autonomous, or that one is functioning well, or that one has friends[4].

(3) A plausible conception of prudential value must be flexible enough to allow for the fact that good lives come in a variety of forms. Non-internalist pluralist theories are not flexible enough, however, and should therefore be rejected[5].

Again, this is an argument which hits some specific non-internalist claims, but it doesn't really refute the most plausible claims of this type. It is true that the most plausible non-internalist pluralist theories are not as flexible as the desire theory, but they all have a certain degree of

---

[4]However, it is not just possible, but also pretty common, to be unaware of the fact that one is *not* autonomous, or that one is functioning badly, or that one doesn't have any real friends.

[5]Cf. the first argument for the desire theory on pp 193-195.

flexibility. Or more specifically, the most plausible non-internalist claims are all consistent with "the fact that the best lives for different people may [on a more specific level] contain quite different ingredients", and the reason for this is that they are all formulated on a high level of generality. For example, *filia*-relationships come in several different forms, and autonomous lives may be very different from each other. But the question still remains: Are the most plausible theories of this type flexible enough? Well, if one accepts Sumner's criterion of flexibility (cf. p 194 above), they are *not* flexible enough, but I can't see why we should accept such a strong requirement (especially the idea that a theory of welfare can not be "descriptively adequate" if it favours companionship over solitude; cf. Sumner (1996), p 18).

(4) It is also possible to argue against non-internalist pluralist theories in a subject-oriented way, i.e. by trying to show that there is no way in which the relevant claims can be supported by a plausible conception of human nature[6], or by trying to show that if we really understand what it is to be human, we will see that no such claims are valid[7]. We will not look at arguments of this type until later, however, viz. when we look at the subject-oriented arguments which has (or can) be given *for* theories of this type. In connection with that, we will not only conduct a critical examination of these attempts at justification, we will also see whether reflections on human nature may even weaken the universalist pluralist's case.

To conclude, there is (as I see it) at least one strong argument against all

---

[6]Where I assume that the relevant evaluative claims are not constituent parts of such a conception.

[7]Or as Sumner (1996) would put it: A list of "intrinsic sources of welfare" (i.e. prudential values) is not a *theory* of welfare. A plausible *theory* of welfare must also give us "a formal account of what it is for something to be such a source", or alternatively put, a (subject-oriented) account of why certain things should be regarded as prudential values. And this is a requirement which no non-internalist pluralist theory can (according to Sumner, that is) satisfy.

Now, it is important to point out that the "objective list theorist" does not have to accept this requirement. Instead, he can adopt the following position: "Why should I have to justify anything? As I see it, my claims are too fundamental to allow for any justification at all. In fact, it seems that Sumner's requirement is really a kind of persuasive definition of the phrase 'theory of welfare', and that it is (just like his requirement of flexibility) especially designed in order to give support to subjectivism". (I owe this observation to Mats Furberg).

non-internalist claims, viz. (1). In my view, this argument constitutes a serious threat against all non-internalist pluralist theories, serious enough to place the burden of proof on the pluralist.

The challenge which the non-internalist pluralist has to meet can be described as follows: He has to show us why we should (so to speak) ignore the heavy objection presented above, or in other words, why we should (in the light of this objection) regard things like autonomy, achievement, or *filia* as universal prudential values. That is, he has to find a way to justify a number of claims of the form "All non-intrinsic facts of type X have nonderivative value for all human beings at all times (where X is not desire-fulfilment)"[8]. In particular, he has to come up with a satisfactory *subject-oriented justification* of such claims (cf. e.g., pp 110-112), i.e. he has to show us what it is about us (our human nature, our constitution) that makes things like autonomy or friendship nonderivatively good for us. Or as Sumner (1996) puts it, he has to give us a formal account of what it is for something to be an "intrinsic source of welfare"[9]. So, let us now look at how the non-internalist pluralist tries to deal with this "problem of justification".

Arguments for the "theory"

These arguments can be divided into two categories, viz. subject-oriented arguments, i.e. arguments which purports to justify the "theory" in a subject-oriented way, and other kinds of arguments (some of which can be regarded as object-oriented). The most important arguments are of the former type, but we will start with the latter type of arguments.

---

[8]This is the type of claim that a *pure* "objective list theorist" has to justify. However, it is worth noting that there are also weaker versions of the "theory"; cf. pp 362-366 below.

[9]Cf. note 7 above. Sumner seems to take it for granted that such an account has to be unified, but there is (as I see it) no reason why we should share this assumption. Instead, we should (at least for the time being) allow for the possibility that the relevant (subject-oriented) justifications may vary from case to case (cf. note 7 on p 288).

"Non-subject-oriented" arguments

There are at least three arguments of this type, viz. (1) the idea that theories of the non-internalist pluralist type have, as a rule, more "critical power" than its competitors, or alternatively, that the theories of prudential value which are best suited for criticizing the prevailing social conditions are all of this type; (2) a number of "atomist arguments"; and (3) Griffin's mixture of "definitional" and "radical" deliberation.

The appeal to the critical power of the "theory"

Theories of prudential value can be used in many different ways, but one of the most important functions is "the critical function"; we want such theories to be useful for criticizing social and cultural arrangements, traditions, practices, and the like. Now, non-internalist pluralist theories are (of course) not the only theories which meet this requirement; hedonism and the desire theory have a certain critical potential as well. It can be argued, however, that this critical potential is not as powerful as in the case of certain kinds of pluralist theories. Consider the case of slavery. Suppose that we think that slavery is a detestable social practice, and that the reason for this is (in part) that it is bad for certain people, viz. the slaves. The question then arises: Why is it bad for (actual and hypothetical) people to be slaves: which theory of prudential value gives us the best explanation of why slavery should be condemned? All the hedonist and the desire theorist can do is this: They can point to the fact that there are (actual) unhappy slaves (or slaves who would prefer to be free), and they can argue that it will be better for future generations (especially for the would-be, hypothetical slaves) to grow up in a society free from slavery. But they can not claim that it would be good for a "happy slave", i.e. a slave who is happy with his lot, and who desires to stay where he is, to be set free. This is something that a non-internalist pluralist can do, however, e.g., he can also appeal to the idea that it may be nonderivatively better for us to live autonomously than to live heterenomously, even if the autonomous life is somewhat unhappier. Therefore, his theory has (in this context) more "critical power" than the hedonistic theory or the desire theory.

So, is this a good "argument"? Well, not if it is regarded as an attempt at objective justification (it is really a "source of appeal" rather than a proper argument), but it may well convince certain kinds of "critical

people" that some kind of non-internalist pluralist theory is best suited for their purposes, and therefore has to be accepted.

Some "atomist" arguments

As an example of an argument of this type, consider Sen's (1992) argument for the view that acting freely and being able to choose have nonderivative value for us: "It is, in general, better for a person to do X and choose it (where choice implies the existence of alternatives) than to do X without having a choice (assuming that "doing X" can be distinguished from "choosing X and doing it"; cf. p 52)[10]. Therefore, being able to choose is directly conducive to well-being".

Is this a good argument? Well, it seems plausible to claim that it is better for a person to do X and choose it than to do X without having a choice, at least if we assume that doing X is a valuable "functioning", or that it is (on a more global level) better for a person to live a certain life if other options are available than it is to live "roughly the same life" where these other options are not available. But even if this is so, we can not conclude that choosing or freedom has intrinsic prudential value. First, the alleged fact that doing X and choosing it is better than just doing it does not imply that it is *intrinsically* better. And even if doing X and choosing it is (in fact) intrinsically better than just doing it, this does not permit us to conclude that choosing is intrinsically valuable. Such a thing can only be concluded if it is assumed that the prudential value of a whole is just the sum of the (prudential) values of its parts: "If the value of the whole 'doing X and choosing it' is bigger than the value of just doing it, then the extra value must be located in the extra part, i.e. in the choosing". This assumption is false, however (cf. p 109 above), and we should therefore reject the argument[11]. But

---

[10]Sen (1992) also tries to convince us of this "principle" (i.e. the idea that doing X and choosing it is better than just doing it) by appealing to the idea that fasting is better than starving. This is a bad argument, however, and the reason for this is that Sen is wrong about what the most important differences between fasting and starving consist in. It is true that fasting may be regarded as "choosing to starve when one does have other options" (p 52), while starving implies that one does not have any other alternatives. But it is also likely that (i) while a starving person wants to eat, a fasting person does not want to eat, and wants to not eat, and (ii) that fasting persons suffer less from their not-eating than starving persons (who wants to eat). And it is mainly because of these two differences that fasting is better (or less bad) than starving.

[11]And since there are, as far as I know, no other arguments for the view that

we may still accept the idea that it is (intrinsically) better to act freely than to act without having a choice, and that a person's level of well-being is (in some way, and to some extent) a function of how free he is.

Sen's argument is not the only argument which is based on the dubious atomistic assumption that the value of a whole is the sum of the values of its parts. Here are two other examples:

Suppose that I am happy because I believe the woman with whom I am love is also in love with me. In this case, it is nonderivatively better for me if my belief is true than if it is illusory, i.e. if she is actually in love with me. But this does not allow us to conclude (there are two possible conclusions here) that it is either nonderivatively good for me to be in contact with reality (that my belief is true), or that actual mutual love is good for me.

It is better to have one's desire for a love affair fulfilled than to have one's desire for degradation fulfilled, even if the latter desire happens to be stronger. But this does not allow us to conclude that it is non-derivatively better for a person to have a love affair than to be degraded.

Griffin's mixture of "definitional" and "radical" deliberation

Let us first note that Griffin's (1986) (and (1996)) attempt to justify his own list of universal prudential values - viz. accomplishment, active (or autonomous) living, understanding, enjoyment, and deep personal relations - is both object-oriented and subject-oriented. Here, we will restrict our attention to the object-oriented part, however; the subject-oriented part (which is a kind of "human nature account") will be discussed at a later point.

Griffin's object-oriented argument can be characterized as follows: First, he gives us an account of "prudential deliberation", or (roughly) of how we should deliberate in order to determine what has prudential value. This "general method" can then be used to criticize and assess a number of specific claims about prudential value. The idea is (of course) that if a "candidate for value status" survives this type of criticism (such a process of deliberation), it can be considered a universal prudential value, and the claim that the thing in question has nonderivative value

---

choosing (or freedom, or the like) has intrinsic value, we should also reject this view.

for us can be regarded as justified.

So, what sort of deliberation goes on (and should go on) in deciding the ends of life? How can (and should) claims about them be criticized and assessed? (cf. Griffin (1986), p 64) Well, Griffin himself conceives of the proper kind of deliberation as a kind of mixture between "definitional" and "radical" deliberation. The definitional deliberation gives rise to proposals of the type "enjoyment is a universal prudential value", and if these proposals are not effectively challenged by radical deliberation, we have a good reason to believe that the proposal is correct.

The definitional part of the deliberation follows naturally from his "informed desire theory" (cf. pp 242-244), i.e. from the idea that "[v]alues are, on the most plausible account of their link to desire, what one would want if one properly appreciated the object of desire" (Griffin (1996), p 35) To see whether a certain object is really good for us, one must have a proper understanding of the nature of the object, and this understanding is (to a considerable extent) "definitional". This definitional part of the deliberation is said to consist of two parts:

> First, I should have to bring into focus the candidate for value status, largely by distinguishing it from other values and from the valueless. Then I should have to decide whether [this possible prudential value] /.../, finally seen plainly, is indeed valuable. This exercise looks like, and in some sense is, a process of discovery, and it looks as if the value discovered is valuable quite apart from my personal desires and inclinations - indeed, is valuable for humans generally (ibid., p 20).

To get a more precise idea of how this "definitional exercise" is supposed to work, let us see how Griffin tries to justify the idea that *accomplishment* is good for us[12]. His reason for picking out accomplishment as a candidate value in the first place is that the "rough, intuitive case for it is plain. *We all want* to do something with our lives, to act in a way that gives them some point or substance" (Griffin (1986), p 64, my italics). That is, his starting point seems to be subjective. But once the candidate for value status has been picked out in this "rough" way, the question arises: What *sort of* accomplishment is it that would have this

[12]In what follows, the two parts of Griffin's argument - the formulation of the "method of deliberation" and the idea that accomplishment is a value which survives the test in question - will not be kept separate.

status? (cf. ibid., p 64). This is where the definitional deliberation really begins. The first step is to distinguish accomplishment "from other values and from the valueless". Here, Griffin claims that accomplishment has to be kept distinct from things like simply wanting one's life to be valuable, bare achievement (to reach the goals one sets oneself), the development and exercise of skills, and the winning of respect and admiration. He also gives a rough (positive) characterization of the valuable kind of accomplishment, e.g., he claims that it is valuable independently of its consequences, that it need not be of something lasting, or of something of wide or public importance. The outcome of this first step of definitional reasoning is that the candidate value (in this case, accomplishment) is "seen plainly". "It [this definitional kind of reasoning] whittles away what only looks like, or is confused with, the end itself. After whittling away, there might be nothing left; or there might be no new value, only an old one now separated from confusing appearances" (ibid., p 65). This is where the second step must be taken, i.e. one then has to "decide whether accomplishment, finally seen plainly, is indeed valuable". Griffin writes:

> But why, once a candidate value is tolerably well defined, accept it as a value? Well, all that one can do to show that it is not a value is to keep whittling away - to show that what makes a candidate value look attractive is something else with which it is confused, or is something meretricious, which when isolated and seen plain is no longer attractive. But once all whittling is done, one must just make up one's mind: is accomplishment, now that it is separated off and seen plain, worth going for? (ibid., p 66).

But, as Griffin points out, "[i]t is not that this is the end to all possible argument. It is just that, once the whittling is done, the argument has to become more radical" (ibid., p 66). An example of such a radical argument is the idea that everything is pointless, that nothing is really worthwhile. If a candidate value cannot survive this type of radical challenge, it cannot be considered a universal prudential value. (And it goes without saying that Griffin believes that the prudential values on his own list would survive such criticism as well)[13].

---

[13]We should also add that there is (on Griffin's view) more to prudential deliberation than what has just been described. A full account of prudential deliberation should, he thinks, also include an account of what the *deliberating*

So, is this a good argument? Well, in my view, the "general method of deliberation" he proposes is plausible enough, but I am not sure whether the method (when applied) "selects" (or generates) the list of prudential values that Griffin thinks it does. For example, a hedonist who is "whittling away" in his own way (rather than in Griffin's way) will probably end up with what he considers a justification of hedonism. And even if the argument happens to give some support to some non-internalist pluralist theories, I doubt that it can counterbalance the "whittling away" that was done in counter-argument (1)[14].

To sum up, none of the three "non-subject-oriented arguments" above is good enough to give us any reason to accept a non-internalist pluralist theory[15]. So, let us now turn to the attempts which have (or can) be made to justify the "theory" in a subject-oriented way.

The subject-oriented arguments

There are at least three subject-oriented ways in which the non-internalist pluralist can try to justify his claims. The three types of subject-oriented justifications are[16]:

(1) *Subjectivist justification*. Here, an attempt is made to justify the relevant universal claims by referring to certain actual attitudes which we all share, e.g., to our intrinsic desires or evaluations. That is, justifications of this type are all "based on" the object interpretation of some actual desire theory, regarded as a theory of (subject-oriented) justification.

(2) *Quasi-subjectivist justifications* are (instead) based on some idea-

---

subject must be like for this type of deliberation to be successful (cf. Griffin (1996), p 58). For example, the deliberating subject must have a well-developed conceptual apparatus that allows him to think about these issues, and he must be sensitive to prudential value. On Griffin's view, this sensitivity is "more judgement-like than perception-like", and an "account of the conditions for the successful working" of this sensitivity will regard these conditions as "akin to conditions such as good light, good eyes, and good position for successful seeing".

[14]We should also add that in the case of accomplishment, counter-argument (2) is also applicable, and the reason for this is that accomplishment is the kind of thing of which one need not be aware.

[15]But cf. the second part of note 7 on p 306 above.

[16]It is important to note that the first two of these "possible justifications" are (probably) my constructions: As far as I know, there have been no actual attempts to justify non-internalist pluralism in any of these ways.

lized version of the desire theory, i.e. they are attempts to justify the relevant claims by referring to what we would want or value under ideal circumstances, e.g., if we were fully rational.

(3) *Objectivist justification*. The objectivist non-internalist pluralist (or "*objective* list theorist", in the proper sense of the term) rejects the idea that things have prudential value in virtue of our attitudes (actual or hypothetical). This means that he can not justify the relevant universal claims by referring to what we want or value, or to what we would want or value under more ideal circumstances. Instead, he has to appeal to something in us which can not be regarded as an attitude or concern, e.g., to a common human potential, or to a set of universal human needs.

Let us now take a closer look at these subject-oriented attempts at justification, to see if any of them are successful.

Subjectivist justification

According to pure subjectivism, the relevant universal claims can (and must) be justified by referring to (intrinsic) desires or evaluations which we all share. As an illustration of this type of justification, consider how a subjectivist would try to justify the idea that it has nonderivative value for all human beings (at all times) to have *filia*-relationships with other people. In this case, he would claim that the reason why it is always nonderivatively good for all of us to have friends is that we all have an intrinsic desire to have friends (or that each of us believes that it is good for him or her to have friends). This argument has the following structure[17]:

(P1; the factual premise): We all have an intrinsic desire for friendship (or every human being believes that it is good for him or her to have friends; or it is "H-evident" to each of us that it is good for him or her to have friends).

(P2; the principle of justification): If we all have an intrinsic desire (etc.) for a certain type of situation, then it has nonderivative value for all of us that situations of this type obtains (this principle can be deduced from the object interpretation of the actual desire theory).

Conclusion: It is nonderivatively good for all of us to have friends,

[17]Cf. pp 113-118 above.

i.e. friendship is a universal prudential value.

So, is this a good argument? No, it is not, and the main reason for this is that the factual premise is false, not just in the case of *filia*, but also (and "even more so") if we have alleged values like autonomy, or contact with reality, in mind. We want different things, and this makes it impossible to justify *universal* claims in this way. Or alternatively put, the main reason why we can not accept the conclusion is that (P2) rests on the object interpretation of the desire theory, a theory which (when combined with the idea that we want different things) implies relativism. This suggests that the best the subjectivist can do is to generate what Sumner (1996) would call "a list of standard intrinsic sources of welfare".

But this is not the only reason why we should reject all subjectivist attempts to justify universal claims. Suppose (for the sake of argument) that both (P1) and (P2) are valid. Now, if (P2) is valid, we can safely assume that the reason for this is that the object interpretation of the desire theory is valid. But if this is so, it is rather pointless to look for universal prudential values at all, isn't it? In this case, we could just say that it is good for me to have friends because *I* desire to have friends, i.e. we could just ignore the issue of whether it is also good for everybody else to have friends[18].

Quasi-subjectivist justification

The quasi-subjectivist universalist tries to justify the relevant claims by referring to what we would want or value under ideal circumstances, e.g., if we were fully rational[19]. For example, he could try to justify the idea that it has nonderivative value for all human beings to function

---

[18]To see how silly it would be for a subjectivist to accept the universalist idea that nothing can be good for me unless it is also good for everybody else (or for a universalist to accept subjectivism), consider the following argument: Suppose that I have two intrinsic desires, viz. the desire for friendship (which is universal), and the desire to be in touch with reality (which is not). In this case, a universalist subjectivist would claim that it is good for me to have friends, but not to be in touch with reality, i.e. he would suggest that what is good for me depends (in part) on what other people want. But this is absurd; what is good and bad for us cannot be determined by consensus decisions (or majority decisions either, for that matter).

[19]Where valuing is, in this case, a special case of desiring, viz. if we assume that a fully rational person will not desire anything he regards as bad, and that he will desire everything which he regards as good.

autonomously by referring to the "fact" that if we were fully rational, then we would all regard autonomous living as a prudential value (at least in one's own case). This argument can be structured as follows:

(P1; the factual premise): If we were fully rational, we would all have an intrinsic desire for autonomous living. Or alternatively put, every fully rational person desires (as an end) to function in an autonomous way.

(P2; the principle of justification): If all rational people have an intrinsic desire for a certain type of situation, then it is nonderivatively good for all of us that situations of this type obtains (this principle can be deduced from the idealized desire theory).

Conclusion: It has nonderivative value for every human being that he or she is functioning in an autonomous way, i.e. autonomous living is a universal prudential value.

So, is this a good argument, i.e. can universal claims about prudential value be justified in this way? For example, is the idealized desire theory on which the second premise is based compatible with the idea that there are universal prudential values, or does it imply relativism? Well, this depends on whether the conception of rational desire that is adopted is "universalist" or "relativist", i.e. whether or not it is implies that "if it is rational for me to desire a certain thing, then it is also rational for you, and vice versa".

If some relativist conception of rational desire is adopted (e.g., the deliberative theory), then the same type of argument that was given against the possibility of subject-oriented justification (of universal claims) can also be given here. That is, (P1) is false, or alternatively, the version of the idealized desire theory (or quasi-subjectivism) on which (P2) is based implies (when combined with the idea that fully rational persons may want different things) relativism.

So, what if some universalist conception of rational desire is (instead) adopted, e.g., the intrinsic theory? Well, in this case, the argument is no longer open to the objection above, but this does not mean that it is a good argument. First, there are good reasons for believing that all idealized desire theories should be rejected (cf. pp 275-278 above), and second, even if we (for the sake of argument) assume that (P2) is valid, it is simply not "fundamental enough" (or "complete enough") to be

satisfactory in this context. Suppose we were to find out that if we were rational, we would all desire to function autonomously. This "fact" would probably give some support to the idea that it is good for us to live our lives in an autonomous way, but it would still be appropriate to ask *why* we would have this desire if we were rational, i.e. what it is about us human beings that explains this alleged fact[20] (cf. pp 119-120). And if this something is not actual desire (e.g., of a very deep kind), we have to look in the direction of objectivism.

Objectivist justification

So, it seems that the non-internalist pluralist can not justify his universal claims by referring to what we want or value, or to what we would want or value under more ideal circumstances. That is, it seems that he has to justify his claims "objectively" after all, i.e. by referring to some aspect of the our nature, other than our subjectivity, some aspect which is not of an "attitudinal" nature. But how can this be done?

The problem can also be characterized as follows: The objectivist non-internalist pluralist (or objective list theorist) claims that there are certain things which are good or bad for us, "whether or not we want to have the good things, or avoid the bad things" (cf. Parfit (1984), p 499), i.e. that "something can be (directly and immediately) good for me though I do not regard it favourably, and [that] my life can be going well despite my failing to have any positive attitude toward it" (Sumner (1996), p 38). But if "personal concerns play no role in determining why something (anything) counts as a good for /.../ [us], or why one thing counts as a greater good than another" (cf. ibid., p 215), and if our well-being does not depend on our attitudes or concerns (actual or hypothetical), then how can we determine what is nonderivatively good and bad for us? And how can claims like "it is good for all of us to achieve things (or to have *filia*-relationships with other people, or to be in contact with reality, or to live our lives autonomously)" be justified? Or as Sumner (1996) puts it: What is it that "makes something (anything) a source of our welfare - what gains it a place on the list - if this does not depend on our attitudes and concerns"? (p 46). This is the challenge that the objectivist non-internalist pluralist (*objective* list

[20]Assuming that it can be explained, that is. Cf. also note 7, the second part.

theorist) has to meet[21].

So, let us now look at some of the ways in which objectivists have actually tried to meet this challenge, to see if any of these attempts are good enough to merit our approval (i.e. to see whether it is possible to justify the relevant universal claims objectively, and if so how)[22].

---

[21]At this point, it is important to point out that even though universalism seems to presuppose objectivism, the reverse is not true, i.e. there are objectivist theories which imply relativism (with respect to what has nonderivative value for us); and just as universal claims of the form "facts of type X is good for all of us" can be given objectivist justifications, so can relativist claims of the form "facts of type X is good for me, but not for you (where X is *not* an instance of anything which is nonderivatively good for everyone)".

An objectivist (subject-oriented) justification of such a relativist claim, e.g., the claim that it has nonderivative value for me to be engaged in physical activity, but not for you, will appeal to some "non-attitudinal" difference between us. Or alternatively put, an "objectivist relativist justification" of the claim that situations of type X are good for a person P is a justification which will refer to aspects (features) of P which are neither shared by all human beings (they may even be unique to P), nor part of P's "subjectivity", e.g., to P's (individual) needs, talents, abilities, callings, vocations, or the like.

Now, it might be argued that the objectivist views on which these justifications are based - e.g., the idea that it is good for a person to have his (individual) needs satisfied, or to act in accordance with his vocation - are really universalist views, and that they (for this reason) do not really imply relativism (with respect to what has prudential value) after all. For example, consider the view that it is good for all human beings to act in accordance with their vocation. Isn't this a universalist view? Well, in a sense, it is, but this does not mean that it implies that the same types of things have nonderivative value for all of us. The reason for this can be formulated as follows: Just as there are two possible interpretations of the idea that it is good for us to have our desires fulfilled (viz. the object interpretation and the satisfaction interpretation), so there are two possible interpretations of the view that it is nonderivatively good for a person to act in accordance with his vocation. On the first interpretation (which corresponds to the satisfaction interpretation of the desire theory), prudential value is attributed to the circumstance that someone acts in accordance with his vocation; and on the second interpretation (or the object interpretation), value is (instead) attributed to the object of a person's vocation, i.e. to the very activity to which he is "called". Now, it is only the first (and less plausible) interpretation (where the idea is understood as a substantive evaluative claim) which implies universalism. On the second interpretation (which is also the more plausible), the idea that it is good for all human beings to live in accordance with their vocation is a formal theory about how we should determine what has value for a person, and if we assume that vocations are something that vary from person to person, it is clear that this theory implies relativism.

[22]And just as it is possible that none of these attempts are successful, we should also allow for the possibility that several attempts are successful. That is, it is (again) important to note (cf. note 7 on p 288 and note 9 on p 307) that "the best objectivist justification" of the relevant universal claims may differ from case to case, and that there may be no (satisfactory) unified objectivist account of what

As far as I can tell, these objectivist attempts are all attempts to justify the relevant universal claims by appealing to a common human nature: they can all be regarded as "human nature accounts". Or alternatively, all objectivist attempts at subject-oriented justification can (it seems) be regarded as instances of the following formula: "The reason why it is nonderivatively good for all human beings to have friends (or to function autonomously, or the like) is that having friends (etc.) is in accordance with human nature (or 'good human nature')".

The differences between objectivists are (roughly) differences with regard to what "non-attitudinal" aspect (or part) of our common human nature they appeal to, and what type of appeal it is (i.e. "how" they appeal to human nature). For example, some objectivist accounts of well-being can be regarded as perfectionist while other must be regarded as "non-perfectionist", some objective accounts appeal to needs while other accounts appeal to the human potential, and so on.

In my view, the most fundamental (and fruitful) distinction between different human nature accounts is the distinction between perfectionist and non-perfectionist accounts. So, how do these two types of accounts differ from each other? What is it that makes a human nature account perfectionist?

Perfectionist accounts (justifications)

Perfectionist justifications are (roughly) justifications which appeal to the human potential or to "human flourishing". So, what kinds of universal evaluative claims is it possible to justify in this way? Well, it seems that there are no claims which can not (in principle) be given perfectionist justifications, but as I see it, it is primarily claims of the form "F is the best (or perfect) way for a human to function (or live, or be)" which can be justified in this way. That is, if other kinds of value-for-claims are to be given perfectionist justifications, this has to be done indirectly, *via* some perfectionist account of human functioning.

So, this is how a claim of the form "F is the best (or perfect) way for a human to function" (e.g., "the best way for a human to function is to function autonomously") can be "justified" in a perfectionist way: "The reason why F is the best way for a human to function is that it is the way of functioning in which human nature flourishes and reaches per-

---

makes things good and bad for us.

fection to the highest degree" (cf. Griffin (1986), p 56)[23]. Or in terms of the human potential: "F is the best way for a human to function because it constitutes a realization of the human potential"[24].

"Justifications" of this type is based on the view that there is such a thing as a "perfect way for a human to function - the spirit in which to approach life, meet its trials, decide on goals and action" (ibid., p 63), a way of functioning "in which human nature flourishes and reaches perfection", or which can be regarded as a realization of the human potential[25]. Now, it is important to see that this idea does not imply that it has *nonderivative* value for us to function in this way. (If this is to follow, we have to add the assumption that it has nonderivative value for us to function in a perfect way, and this assumption is not entirely unproblematic).

But more importantly, as it stands, this perfectionist account of good human functioning cannot be used to justify any other claims of prudential value (i.e. besides the claim that it has nonderivative value for us to function in a perfect way). So the question arises: If the perfectionist account of human functioning is to be used to determine what it is besides perfect human functioning that has universal prudential value, then how should this be done? For example, what extra assumptions have to be made?

Well, let us first note that if the relevant universal claims (e.g., the claim that it is good for all of us to have friends) are to be justified in a perfectionist way, we have to regard the perfectionist account above as more that just a substantive account of ideal human functioning, viz. we

---

[23]But is this really a type of *justification*? Isn't the idea better expressed as "F is the best way for a human to function *if and only if* it is the way of functioning in which human nature flourishes (etc.)"? Well, on my view, the idea is best expressed as "F is the best way for a human to function *if and only if (and because)* it is the way of functioning in which human nature flourishes (etc.)", and it is the presence of the term "because" that makes it appropriate to conceive of the idea as some kind of "justification".

[24]It worth pointing out that this type of perfectionism need not be monist: one may also be a pluralist in this area, and claim that F1 is one of several perfect ways for a human to function *because* it is one of the ways "in which human nature flourishes and reaches perfection". Such an idea is probably based on the assumption that there are several different "kinds of human nature" (e.g., innate personality types, or the like), each of which can be more or less "flourishing".

[25]This view may also (but need not) include the idea that the level of well-functioning "for any person is in direct proportion to how near that person's life gets to this ideal".

also have to regard it as an essential part of a "formal account of the modes of approach that will fix on" what (if anything) that has non-derivative value for all human beings (cf. ibid., p 63). This is an example of what such a formal account might look like: "If we would all 'have' or 'achieve' situations of type X if we were functioning in way F under standard (or acceptable) circumstances, then situations of type X are good for all of us; and the reason why X is good for us is that X is an essential part of the kind of life that we would have if we were functioning in the way F under normal circumstances"[26], where we should add (i) that F is the best way for a human to function, and that the reason for this is that F is the way of functioning in which human nature flourishes and reaches perfection to the highest degree; (ii) that there is no reference to X in the specification of F; and (iii) that the notion of standard or acceptable circumstances is also specified independently of X.

Now, formal accounts of this type are obviously problematic, and it is, for this reason, unlikely that they can be used to justify any other universal claims of prudential value besides the claim that it has non-derivative value for us to function in a perfect way[27]. My tentative

---

[26]The reason why I put it this way is that I think it would be too strong to claim that X is good for us *if and only if* (and because) we would all "achieve" X if we were functioning in way F under normal circumstances.

It is important that this "appeal to the well-functioning person" is distinguished from another (similar) idea, viz. the idea that we can determine what is good for us by way of the *attitudes* of the well-functioning person (or "good man"). To see why this idea has no relevance whatsoever in the present context of justification, consider the following two ways of making it more precise: (i) "X is good for us if and only if (and because) well-functioning persons desire, like, or value X", and (ii) "Well-functioning persons like (etc.) what is good for them, and they like these things because they are good for them, and we can therefore use the attitudes of the well-functioning man as an indication of what is good for all of us". Now, (i) is clearly a justificatory idea, but if we do not view it as just another formulation of quasi-subjectivism, it is surely an absurd claim (it is absurd to claim that something is good for me - who is not well-functioning - because someone else values it; cf. note 18). And while (ii) is a more plausible idea, it is totally irrelevant in the context of justification. But to see why (ii) is probably implausible too, consider some of the assumptions on which it is based, viz. (a) that there is such a thing as "what the good man values (etc.)"; (b) the universalist view that what is good and bad for one person (e.g., a "good man") is also good or bad for all of us; and (iii) that the reason why the good man is correct about what is good and bad for him (and everybody else) is that he is (because of his goodness) more in touch with our common human nature than the rest of us.

[27]There are several reasons why these accounts have to be regarded as problematic, e.g., (i) it seems implausible to assume that we can specify both good

conclusion is therefore that a universal claim of prudential value cannot be given a perfectionist justification unless it is of the form "it has non-derivative value for all of us to function in way F (e.g., autonomously)".

The only perfectionist argument that we will take a closer look at, viz. Aristotle's famous *ergon* argument, is of this type, i.e. it is an attempt to justify the claim that it has nonderivative value for all of us to function in a certain way, viz. "in accordance with excellence" (cf. pp 298-301 above). The argument appeals to the *ergon* of man (which plays an essential role in Aristotle's teleological view on human nature), and it has (roughly) the following structure: "It is good for all human beings to function in accordance with excellence *because* this is what it is the business of a human being to do; this is our characteristic 'activity', this is what makes us human".

"Non-perfectionist" human nature accounts (justifications)

On the non-perfectionist views, it is of no importance what (if anything) makes human nature flourish and reach perfection to the highest degree, or what (if anything) constitutes a realization of the human potential; a non-perfectionist may even reject the assumption that there is such a thing as "reaching perfection to the highest degree", or "realizing the human potential". Instead of appealing to these (possibly empty) notions, non-perfectionists appeal to other aspects of our common human nature, and they do it in a very different way.

The non-perfectionist objectivist justifications which we will look at are of three different kinds, and so are (of course) the human nature accounts on which these justifications are based:

(1) *The basic need account*. Here, the relevant universal claims are justified in terms of universal human needs, i.e. in terms of what we need *qua* human beings. These justifications have the following form: "The reason why X is a universal prudential value is that X is something we

---

functioning and good-enough circumstances independently of any (substantive) conception of prudential value, and (ii) it also seems implausible to assume that "good functioning under good-enough circumstances" will, *in all cases*, give rise to lives which are similar in content. Now, (ii) suggests that we could improve our case if we exchange the universalist account above for the following kind of relativist account: "The reason why X is good for a person P is that X is an essential part of the kind of life that *P* would have if he were functioning in the way F, where F is the perfect way for a human to function (or for *him* to function), and so on". It is doubtful whether this account can escape objection (i), however.

all need *and* this need is a basic human need".

It seems that all kinds of universal claims can (in principle) be justified in this way, but it surely sounds more odd to talk about a basic need for autonomous living than to talk about the need for friendship or love.

(2) *The appeal to what is and what is not "recognizably human".* Justifications of this type have the following form: "The reason why X is a universal prudential value is that a life (or existence) which does not contain X is not recognizably human". A good example of such an argument is the Aristotelian idea that it is good for all human beings to have friends because human beings are social beings, or "political creatures" (where it is assumed that it is not part of the *ergon* of man to be a social being).

It seems that all kinds of prudential values can (in principle) be "grounded" in this way, but it is important to note that the kinds of universal claims that are best justified in this way are claims to the effect that it is good for us to have a certain minimal amount of certain goods, e.g., ideas like "the more autonomous a person is, the better" can hardly be justified in this way.

(3) Griffin's human nature account, where an appeal is made to a number of different *"aims" and "interests"* which are (supposedly) *embedded in human nature.*

This concludes our brief list of possible objective (subject-oriented) justifications. So, let us now take a closer look at some of these alleged justifications (and the views on which they are based), to see if any of them are plausible enough to give us a good reason to accept some non-internalist pluralist theory. To repeat, the accounts which will be discussed are: (1) Aristotle's *ergon* argument; (2) the basic need account; (3) one of several possible appeals to what is and what is not "recognizably human", viz. the Aristotelian attempt to justify the idea that it is good for all human beings to have friends by referring to the "fact" that human beings are social beings, or "political creatures"; and (4) Griffin's appeal to "aims" and "interests" which are embedded in human nature.

A critical discussion of four different human nature accounts

The first Aristotelian appeal to Human Nature: the *ergon* argument

The general assumption on which the *ergon* argument is based is the idea that "[f]or all things that have a function or activity [ergon], the good and the 'well' is thought to reside in the function [ergon]" (*NE*, I.7, p 13). Or in Nagel's (1972) terms, "if something has an ergon, that thing's good is a function of its ergon", or alternatively, "when something has an ergon, that thing's good is specified by it" (p 8). That is, if we know what a thing's ergon is, then we can draw certain evaluative conclusions. But what kind of evaluative statements is it possible to derive from ergon-statements?

To find an answer to this question, we must first know more about what the ergon of a (kind of) thing is. According to McDowell (1980), the ergon of F is "what it is the business of an F to do"; according to Nagel (1972), "[t]he ergon of a thing, in general, is what it does that makes it what it is" (p 8); and according to Wilkes (1978),

> [t]he ergon of any X is the function that it has; or, if it is the kind of thing that cannot readily be said to have a function, it is its characteristic activity. It is definitionally assigned; it is what X does that makes it just what it is, and if for any reason X becomes unable to perform its ergon, it is then no longer genuinely an X at all (p 343).

Now, this suggests that if we know what the ergon of K (a *kind* of thing) is, then what we can conclude is what a good (or excellent) K is. And if we know what a good K is, we can also determine whether a particular thing X (of the kind K) is a *good of its kind*, i.e. whether it is, *as a K, good*. That is, the conclusion of an "ergon argument" is (it seems) always of the form "a good K is a K which is constituted in this-or-that way", and *not* (as e.g. Nagel seems to suggest) "this is what is good for a K".

To see what kind (or form) of goodness that is involved in conclusions of ergon arguments, we have to consider the fact that in this context, goodness is attributed to things which are what they are in virtue of having certain functions or characteristic activities. This suggests that

the forms of goodness that a thing has in virtue of its ergon is either *instrumental goodness* or *technical goodness* (the terminology is from von Wright (1963)). If X is an instrument or a tool (e.g., a knife), then its ergon is some function that it has which makes it what it is (i.e. some purpose which is "essentially associated with the kind" to which it belongs), and we can say that it is (instrumentally) good if it serves this purpose well. And if X is, for example, a chess-player, then his ergon (*qua* chess-player) is the activity of playing chess, and we can say that he is a (technically) good chess-player if he is good at this activity. If this is correct, we can (it seems) conclude that the conclusion of an ergon argument is either of the form "an instrumentally good K is a K which serves this-or-that purpose well", or of the form "a technically good K is a K who is good at this-or-that activity".

Now, if we turn to the *ergon* argument itself, the first thing that strikes us here is that the conclusion of this argument is *not* a claim about what the good (or well-functioning) man is like, but a claim about what is *good for man* (or what "the good life for man" consists in). Or more specifically, the conclusion of the argument is the idea that it is (nonderivatively) good for us to function in accordance with excellence[28], and this conclusion is (somehow) thought to follow from a statement about what the ergon of man is.

So, what is the ergon of man? What is it "the business of a human being to do", and what is it that makes him a human being rather than something else? On Aristotle's view, the ergon of man is "an activity of soul which follows or implies a rational principle" (*NE*, I.7, p 13), or "activity of the *psuche* in accordance with a rational principle". To grasp the meaning of this idea, we have to take his distinction between two major forms of rationality (which is based on the way in which he divides the soul) into account, viz. the distinction between practical wisdom (*phronesis*) and philosophic wisdom (*sophia*). So, as Wilkes (1978) points out, "man's ergon may be the activity of the psuche in accordance with either or both of these" (p 343). In short, the ergon of man consists in rational activity, or functioning in accordance with theoretical

---

[28]But it is important to note that what Aristotle *really* argues for is a much stronger claim, viz. the view that *eudaimonia* (or "the good life for man") consists in "activity in accordance with excellence". On this view, functioning in accordance with excellence is the "chief human good", and not merely one of several prudential values.

and/or practical reason[29].

Now, it is clear that all that can be concluded from

(P1) The ergon of man (assuming that there is such a thing) consists in rational activity (functioning)

and the fundamental assumption that

(P2) For all things that have an ergon, the good and the 'well' (i.e. its technical or instrumental goodness) is thought to reside in the ergon"

is the following "lemma":

(L1) To be a good man, to function well as a human being, is to function in a rational way.

But if we assume that Aristotle's list of excellences (of intellect and of character) is a good specification of what it is to function in a rational way, i.e. that

(P3) Rational living consists in activity in accordance with the excellences

then we can also conclude that

(L2) To be a good (well-functioning) man is to function in accordance with the excellences, i.e. good human functioning consists in activity in accordance with the excellences[30].

What we can not yet conclude is what Aristotle wants us to conclude,

---

[29]Some commentators, e.g., Wilkes (1978), wants to include more than rational activity in the ergon of man. Wilkes writes: "Man is a hybrid and two-sided creature, who shares properties with both gods and animals. The ergon offered by Aristotle highlights the side of rationality at the expense of animality and thus oversimplifies the nature of man. With this oversimplification the gap between the life of a good man and the good life for a man appears to widen yet further" (p 345). The problem with this interpretation is that it fails to take into account that the ergon of man is only one part (or aspect) of human nature. As we will see below (on pp 340-345), Aristotle does take our "animality" into account, but in a different context, and this means that we need not agree with Nagel (1972), when he claims that reason is (*qua* ergon idion of man) what human life is about, and that the intellect should (therefore) be included in the account of what a human being is, while the bodily functions which we share with animals and plants (e.g., nutrition) can be excluded.

[30]On this view, a good man is someone who is *good at* living in accordance with reason, and the goodness of the good man should therefore be regarded as a kind of technical goodness.

viz. that

(C) It has nonderivative value for us to function in accordance with excellence.

For this claim to follow, we also have to add yet another assumption, viz.

(P4) It is nonderivatively good for us to function well[31].

To determine whether this is a good argument, we should take a closer look at its weakest point, viz. (P1)[32]. So, is (P1) a plausible assumption? Well, I think not. First, it is doubtful whether there is such a thing as the ergon (idion) of man, i.e. whether there is a kind of *activity* or function in virtue of which we are human. (To doubt that there is an ergon of man is not necessarily to doubt that there is a human nature, however; it is only to doubt that this nature can be defined in terms of activity or function). And second, even if there happens to be such a thing as the ergon of man, it may be doubted that it consists in rational activity (e.g., it may be argued that the ergon of man is, instead, play or creative activity). So in my view, we have little or no reason to believe both that there is an ergon of man *and* that it consists in rational activity, and we should therefore reject the *ergon* argument as a whole[33].

---

[31]If we have Aristotle's "original" conclusion in mind (cf. note 28), this premise can be (roughly) formulated as follows: "The good life for man is the same thing as the life of the good man". It is probably this idea that Sumner (1996) has in mind when he criticizes Aristotle for "reducing" prudential value to perfectionist value (or for "conflating" the two values with each other).

[32]That is, I assume that the conclusion follows from the premises, and that (P2), (P3), and (P4) are all more plausible than (P1). This does not mean that the other three assumptions are plausible, however. For example, many philosophers would probably reject (P4), an assumption which I personally happen to regard as quite plausible (I can't see how one could argue for or against this assumption, however).

[33]It can also be doubted whether it is really possible to conduct value-free investigations into the ergon of man, and whether claims about erga are descriptive claims (especially in the case of living organisms). That is, it may well be the case that even if the *ergon* argument were sound, it would not constitute any bridge between "the is" and "the good".

The basic need account

To find out whether the relevant value-for-claims can be justified in terms of basic human needs[34], we must first know what a basic need is. So, what kind of human needs are basic needs, and how do basic needs differ from other human needs of the same genus? My answer to this question can be formulated as follows: (i) Basic needs are "teleological needs" rather than "tension-needs"; (ii) basic human needs are (typically) universal needs, i.e. they are shared by all human beings, and the reason for this is that (iii) basic needs are (at least in part) rooted in some objective (i.e. non-attitudinal) aspect of our common human nature: basic needs are never dependent on desires, they are never derived from our attitudes and concerns (this is of course the feature of the basic need account that makes it objective); (iv) the goal of a basic need is always valuable, but any kind of valuable situation does not count as a goal of basic need; only a limited number of valuable situations may count as goals of basic need; (v) it is (by definition) good for us to have our basic needs satisfied, and it is (typically) bad for us not to have our basic needs satisfied (this point is of course closely related to the fact that the goals of our basic needs are always valuable).

Let us now take a closer look at these points.

(i) Basic needs are teleological needs

When we attribute needs to people, we either have "tension-needs" or "teleological needs" in mind. A *tension-need* is a state of tension (or disequlibrium) that an organism may be in - a state that functions as a motivating force -, and the object of such a tension-need is the thing that the tension forces the individual towards. Liss (1990) defines "tension-need" as follows: "P has a [tension-] need /.../ for X if and

---

[34]It is important to note that such justifications might, but need not, be based on the "pure basic need account", i.e. on the view that a person's well-being is a function of to what extent his basic needs are satisfied. (Cf. Griffin (1986), according to whom the basic need account is the view that "well-being is the level to which basic needs are met so long as they retain importance" (p 52)). A basic need account need not be pure, however, i.e. it may well allow for the possibility that there are prudential values which are not good for us in virtue of being needed, and it may (as far as I can see) also allow for the possibility that there are basic needs whose satisfaction does not contribute directly to the well-being of the "needing subject".

only if, a) P has a tension, and b) that tension disposes P to X" (p 58)[35]. Now, it is rather obvious that such tension-needs are best conceived of as desires; to have a tension-need for something is very much like having a craving for it, and cravings are properly regarded as desires (in this case, *desires-for* rather than *desires-that*). That is, all attempts to justify goodness-for-statements in terms of tension-needs are (for this reason) subjectivist rather than objectivist, and we can therefore (in this "context of objectivist justification") do without the notion of tension-need altogether.

In this context, the relevant notion of need is *teleological need*; needs of this kind are *not* desires, and they may have an essential role to play in objectivist justifications of goodness-for-statements. Now, as I see it, it is (in this context) not really necessary to know what a teleological need *is* (e.g., what kind of thing it is); it is sufficient to restrict our attention to *statements of need* (to their structure, truth conditions, and the like). All need-statements are of the form (or can, at least, be expressed in the form) "a subject S (in this case some individual human being) needs an object X in order to achieve the goal Y (under the circumstances C)", where the object of need is (typically, but not always) necessary for the goal of need[36].

There seem to be few restrictions on what the goal of human need can be, and even less restrictions on what can be needed; a person may need cool nerves to plant bombs, for example, or a certain education in order to be get a certain job. But because of the nature of the "in-order-to"-relation, the object of need is always (to a considerable extent) determined by the goal of need, and this means that if a certain goal is given, this puts strong restrictions on what the object of need can be.

---

[35]The problem with this definition is that it presupposes that the object of need is a piece of behaviour, or an activity; for example, we can (on this definition) say that a thirsty person has a tension-need for drinking, but not that he has a tension-need for water (a person can hardly be disposed to water). On my view, the definition would be improved if (b) were replaced by (b') "that tension disposes P to strive for (search for, or the like) X".

[36]Now, statements of basic need are normally of the form "we all have a need for X" or "we all need X" (e.g., "we all need food"), and it may therefore *seem* as if basic needs do not have goals. Statements of this form are always elliptical, however. Basic needs always have goals (e.g., humans need food *in order to survive*); it is just that these goals are often taken for granted, and the phrase "in order to Y" is (therefore) not filled in.

All that has been said about the "in-order-to"-relation (the relation between the object of need and the goal of need) this far is that the object of need is (typically) necessary for the goal of need. Now, there are at least two different ways in which the object (X) can be necessary for the goal (Y): X is either causally necessary for Y[37], or it is a necessary constituent part of Y. That is, if X is necessary for Y, then the relation that holds between X and Y is either a causal relation (a kind of means-end relation) or a part-whole relation (a kind of intrinsic relation, e.g., as when we say that human beings need satisfying human relations in order to have a good life).

This is not the whole truth about the "in-order-to"-relation, however. Consider the following example: Let us assume that we have a need for personal growth (as Allardt (1993) claims), and that the goal of this need is "living well". Does this imply that personal development is *necessary* for living well? I think not. On my view, all Allardt's idea implies is the more modest claim that personal development is something which (normally) *contributes* to a good life, or which *helps* to bring a good life about (where the contribution is of a part-whole type rather than of a means-end type). That is, it seems perfectly consistent to accept the idea that we need personal development in order to live well and (at the same time) reject the idea personal growth is necessary for living well (i.e. that it is a necessary constituent part of a good life).

If this is correct, the relation that holds between the object of need and the goal of need may be of four different kinds: The object may be causally necessary for the goal; it may be a necessary constituent part of the goal; but it may also give a causal contribution to the goal (be a non-necessary means to Y, e.g., help make Y possible); or it may be a "contributive constituent part" of the goal. That is, both the part-whole model and the means-end model have to be stretched to include relations that are not "being-necessary-for"-relations[38].

Another important feature of teleological need is this: The objects of such needs are always situations rather than things (and this holds for

---

[37]Notice that a very long time may pass between "the cause" and "the effect", e.g., as in "all humans need a lot of parental love as infants in order to function well as adults".

[38]Cf. Griffin's (1986) idea that "the means-end model has to be stretched beyond simple cause-effect relations to fit all the cases to which it is usually applied" (note 7, p 327).

the goals of teleological need as well). The reason why "the situation-view" is superior to "the thing-view" is twofold: First, all the different relations which may hold between the object of need and the goal of need (i.e. all four "in-order-to"-relations) are relations which hold between situations rather than between "things" (e.g., things can not be causally necessary conditions, but situations (facts) can). And second, the idea that needs can be satisfied, and that we can understand what it is for a need to be satisfied (as well as determine whether a certain need is satisfied or not), seems to presuppose "the situation-view". It is always a situation (rather than a thing) which constitutes the satisfaction of a need, i.e. if we think of needs as "situation-needs", we always know what it is for a need to be satisfied (since the object of the need is, in this case, identical with the situation which would constitute its satisfaction). We do not always understand what it is for a thing-need to be satisfied, however. In fact, it is only possible to determine whether a "thing-need" is satisfied if the need in question corresponds to one or several "situation-needs". For these reasons, we should accept the idea that when a person needs something, what he really needs is that something is the case (cf. also pp 165-169 above).

This ends our characterization of what it is to need something in the teleological sense[39].

---

[39]But it might still be asked what a teleological need really *is*. We know that it is not (like tension-need or desire) a state that the needing organism is in. Neither is the need itself identical with the object of need (even though the term "need" is sometimes used to refer to the object of need). So what is it? Liss (1990) suggests that a need is constituted by the difference between the actual state that the subject is in and the goal of need (i.e. in relation to the goal, the needing subject can be said to be in a state of lack or deficiency). This is how Liss defines teleological need (on p 52): "P has a [teleological] need [or "difference-need"] /.../ for X in situation S at T if, and only if, a) there is a difference between an actual state of P in S at T and a goal G in S at T, and b) X in S at T is a necessary condition for G". (He also adds, a little later, that the actual state and the goal must be "commensurable"). The problem with this definition is that it implies that all needs are unsatisfied, an idea that I find (especially in an evaluative and justificatory context like this) unacceptable; to have a need is not necessarily to be *in need*. So, what is my own view on the issue? Well, either we can say that there are no such things as needs, or we can say that if there are needs, then they must be regarded as complex relational phenomena that are structured just like need-statements. On this latter view, needs *are* 3-ary relations, and to be "in need" of something is a 4-ary relation (which also includes, as an element, the actual state that the subject is in).

(ii) Basic human needs are (typically) universal needs

That a human need is universal means (roughly) that its *object* is something which is needed by all human beings, and that the reason for this is (in part) that the *goal* of the need is (so to speak) shared by everyone. So, are our basic needs universal in this sense? Well, it is quite clear that the goals of my basic needs (e.g., survival) are identical with the goals of your basic needs, but this does not imply that the objects of our basic needs are the same. Or more specifically: To the extent that we are alike in our personal characteristics (and live under similar circumstances), so are the objects of our basic needs, e.g., we all need (in virtue of our shared biological make-up) nutrition in order to survive, and sleep in order to function properly. But if we are different in some relevant respect, this means that the objects of our basic needs may also be different[40]. A person who is paralysed from the waist down, for example, needs (because of his deviation from the biological norm) a wheel chair in order to move around on his own. That is, if we count his need for a wheel chair as basic (as I think we should), then we must conclude that the object of a basic need need not be universal after all. However, the basic needs which are of interest to us in this context of justification are (as we will see) all universal.

(iii) Basic needs are needs which "flow" from human nature

Many human needs are needs we have because of the ends we happen to choose: it is often the case that when a person needs something, he needs it in order to satisfy some desire that he has (in which case the goal of the need is identical with the object of the desire). Now, basic needs are never dependent on desires in this way, they do not, "like hypothetical imperatives, depend upon the adoption of some purpose /.../ [,] they do not depend upon ends that we just happen to adopt" (Griffin (1986), p 42). This is why we have to regard the basic need account as objective; it "makes well-being independent of desires" (ibid., p 32), its "standard rests on aims flowing from human nature and not on any flowing from a person's particular tastes, attitudes, or interests" (ibid., p 53).

---

[40]This (the fact that we are different and live under different circumstances) is but one reason why we need different things. Another reason (which is not applicable to *basic* needs) why our objects of need may be different is (of course) that our goals of need may be different, e.g., because we want different things.

That is, the reason why basic needs are "objectively given" is that they "flow" from human nature: basic needs are needs we have in virtue of being human. But what is this supposed to mean? Well, it hardly means that the possible *goals* of our basic needs (e.g., survival, health, proper functioning, or good living) are "derived" from human nature. This suggests that it is the *objects* of our basic needs which "flow" from human nature. Is this a plausible claim? Well, we already know that the objects of a person's needs are determined by at least two kinds of things, viz. (i) the goals of need and (ii) the personal characteristics of the needing subject (and the circumstances under which he lives). Now, this suggests that if the goals of need are given, then the objects of need can be derived from (a correct conception of) human nature, but only if we have universal needs in mind. For example, the fact that we all need nutrition, air, and water in order to survive can be regarded as dependent on our human nature, and so can the fact that we need rest in order to be healthy, and that we need human relationships in order to avoid misery.

This does not mean that the line between basic and non-basic needs can be fixed solely in terms of human nature, however. In fact, it seems that the objects of our basic needs are not just determined by human nature, but also by convention. For example, the fact that we seem to need things like availability of medical aid, formal education, interesting work, relationships with work-mates, and opportunities for leisure-time activities in order to achieve some vaguely formulated goal like "living well" is not dependent merely on human nature, but also on the fact that we live in a certain kind of society. But just how is the line between basic and non-basic needs dependent on convention, i.e. where exactly does convention enter the picture? Well, it seems to me that it is mainly the goals of basic need, and/or the specification of these goals[41], that are (to some extent) determined by convention; once these goals are given, it doesn't seem that we have to appeal to convention in order to determine what the objects of basic need are. But regardless of where exactly convention enters the picture; it is worth noting that the mere fact that "the line between basic and non-basic needs may change as society changes" (cf. ibid., p 44) seems to make basic needs (partially,

[41]Cf. Griffin's idea that key terms like "health", "proper functioning", and "harm" "have to be given fresh interpretations in each social setting" (ibid., p 45).

and in some way) dependent on "attitudes and concerns", and this might be sufficient to constitute a threat to the objectivity of the basic need account.

(iv) The goals of our basic needs are always valuable

What are the possible goals of our basic needs? Well, the fact that the objects of our basic needs are (to a considerable extent) determined by these goals suggests that we can start from a list of basic needs, and then infer (from this list) what the goals of our basic needs must be. So, what do we need in this basic way, and in order to what? These are some of the objects of basic need that have been suggested: The most indisputable kind of basic needs are the elementary biological needs, e.g., the need for food (or nutrition), air, water, and sleep (or rest). But "basic need theorists" are never content with this: they always tend to include a number of "psychological" or "social" needs on their lists of basic needs, like the need for warm human relationships, meaningful activity, or variation in life. To see how different the objects of basic need may be, consider Allardt's (1993) claim that things like availability of medical aid, formal education, relationships with work-mates, attachment to family and kin, and opportunities for leisure-time activities are all objects of basic need (cf. pp 89-91). He also regards "the need to relate to other people and to form social identities", "the need for personal growth", and "the need for integration into society and to live in harmony with nature" as basic needs. So, what are the goals that "correspond to" these objects of basic need? Well, some of the objects (like nutrition, air, and water) are things we need *in order to survive*; other things (like rest) we need *in order to be healthy, or to function properly*; and other things we need *in order to "avoid misery, relate to other people, and avoid alienation"* (ibid., p 89, my italics). It has also been suggested that a life in accordance with our good nature, the fulfilment of one's plan of life, the realization of one's vital goals, a decent state of living, self-realization, a meaningful and individualized life, and good living are possible goals of basic need[42].

---

[42]At this point, we can note that even if survival and health are necessary (i.e. in this sense "needed") for a good or meaningful life, it sounds rather odd to say that we need to survive or that we need to be healthy. This suggests that the goals of basic need are types of situations which can not (in turn) be regarded as objects of need. Almost all the candidates listed above share this feature, with the possible

In short, it seems that all the possible goals of basic need can either be regarded as *achievements of something (positively) good*, or as *avoidances of something bad*, or as both[43]. In short, all goals of basic need are (it seems) valuable for us[44], but they need not have nonderivative value for us.

Now, the question is whether all the candidates listed should be regarded as goals of basic need, or whether only some of them should. It has been suggested that the goal of our basic needs is always *the avoidance of harm*, or more specifically, that a need is basic if and only if its object is necessary (etc.) in order to avoid harm, e.g., if it is (as Sumner (1996) puts it) a need for something the lack of which will damage or impair one's life[45].

Now, this is not an implausible idea, but I still think it is "too strong". It is true that the best candidates to the title "a goal of basic need" (e.g., survival, health, proper functioning, and the avoidance of misery) can all be regarded as "avoidances of harm", but this does not mean that whenever the goal of a certain need is the avoidance of harm, then this need should be regarded as basic, and neither does it mean that the goal of every single basic need can be regarded as some kind of avoidance of harm. For example, it is doubtful whether the fact that

---

exception of "relating to other people" (which is probably not sufficiently general).

[43]For example, survival is the same thing as the "avoidance" of death, having a decent life may be regarded as the same thing as avoiding a bad life, and health may also be regarded as a "privative notion".

[44]This follows from the fact that the absence of something which is bad for a person is always good for this person, or more specifically, that the absence of the bad is always derivatively (e.g., instrumentally) good in that it makes a number of good things possible. The absence of the good is not necessarily bad, however, i.e. there is a certain "asymmetry" between good and bad. However, if we restrict our attention to *nonderivative value*, we can see that there is a symmetry between good and bad: Here, it is both the case that (i) the absence of a nonderivatively good state is not necessarily nonderivatively bad (but it is necessarily worse), and (ii) the absence of a nonderivatively bad state is not necessarily nonderivatively good (but it is necessarily better). The reason for this is that between the nonderivatively good and the nonderivatively bad, there is the "nonderivatively neutral"; or alternatively put, the terms "good" and "bad" are (in this case) contraries rather than contradictories.

[45]It is worth noting that this view is almost identical with an idea which has been put forward by von Wright (e.g., in von Wright (1986), Ch. XII, section 6) and others, viz. the idea that a (basic) need is something that it is (by definition) bad for a person not to have satisfied (cf. also pp 336-338 below). If my need for X is basic, then this idea implies that not-X is bad for me, and a good explanation of this is that X is necessary (etc.) for the avoidance of harm, and that harm is (by definition) always bad for me. That is, the reason why it is bad for me not to have a (basic) need for X satisfied is that without X, I will be harmed.

some of us need a lot of nicotine (or coffee, or alcohol) in order to avoid misery makes these needs basic[46], and it is also doubtful whether the fact that the goal of a certain universal need is a good or meaningful life implies that this need is not really basic.

In short, I tend to reject the idea that the goal of every basic need is the avoidance of harm[47]. Instead, I propose the more liberal view that there are several possible goals of basic need, some of which can be regarded as the achievement of something positively good. (But as a rule of thumb, we can say that if the goal of a certain need can not be conceived of as the avoidance of harm, then it is likely that the need in question is *not* a basic need). To make an exact specification of what the goal of a need must be like if this need is to count as basic would not be true to a phenomenon that is best left a little open. There are tough cases, which we can not get rid of by making stipulations. The following questions (formulated by Griffin (1986)) can serve as an example of how difficult it is to determine whether something is a basic need or not: "Is interesting work a basic need? Well, without it, alienation, a kind of social pathology, results. Is education a basic need? Without it, one's intellect will atrophy. And how much education?" (p 43)[48].

(v) The value of basic need satisfaction

Now that we have some idea about what a basic need is, we can see how empty the claim that it is good for us to have our basic needs satisfied really is[49]. The notion of basic need is a value-laden notion, and

---

[46]This is of course an objection which the "avoidance-of-harm" theorist can meet by stipulating what is meant by "harm" in this context, and how much harm that should be avoided (and in what way) if this avoidance is to be counted as the satisfaction of a basic need.

[47]This means that I also tend to reject the idea in note 45, viz. the idea that a (basic) need is something that it is (by definition) bad for a person not to have satisfied.

[48]If someone happens to believe that the answers to these questions are obviously "no", it is likely that he has failed to appreciate that questions of basic need are (in part) *political* questions.

[49]Here, one may ask exactly how the claim that it is good for a person to have his basic needs satisfied should be understood. Is it the circumstance that a basic need is satisfied that is good for the needing person, or is it the object of the need? Well, the latter interpretation (i.e. "the object interpretation") is (I think) the more common one, and it is also better than the former interpretation (i.e. "the satisfaction interpretation"). The reason for this is twofold: First, we tend to think of need accounts as *pluralist* theories of prudential value, and this is only possible if the object interpretation is adopted. And secondly, the object interpretation is more in line with the type of objectivist justifications that we are interested in

the reason for this is (of course) that the goal of a basic need is (by definition) good for the needing subject (cf. pp 335-336 above). So, what does this imply about the value of basic need satisfaction?

Well, we know that the object of a basic need is either a means to the goal of the need (necessary or "contributive"), or it is a constitutive part of the goal (necessary or "contributive"). Now, if the object of need is a *means* to the goal of need, it is (obviously) good for the needing subject to have the need satisfied, but only derivatively good. The reason why this is so is that in cases like this, the object of need (*qua* object of need) *derives* its value from the value of the goal of need. But what if the object of need is (instead) a *constitutive part* of the goal of need? Well, in cases like this it seems that the object of need (*qua* object of need) need not derive its value from the value of the goal of need. For example, if having friends is a constitutive part of a good life, and if it is also intrinsically (rather than "contributively") good, then it is (it seems) the value of the whole which is (in part) derived from (in the sense "a function of") the value of having friends, rather than vice versa. In any case, it seems plausible to make the following assumption: If the object of need is a part of the goal of need, and if the goal of need is merely instrumentally good, then the object of need is also instrumentally good. But if the goal of need is intrinsically good for the needing subject, then the object of need is (on the assumption that it is a constituent part of the goal of need) either intrinsically good or "contributively good" for this subject.

The first conclusion we can draw from all this is that it is always good for us to have our basic needs satisfied. However, it can not be concluded that it is always bad for us to not have our basic needs satisfied[50].

---

here, i.e. justifications of the form "this is good for him *because* he has a basic need for it".

[50]This follows from the fact that the absence of something good is not necessarily bad (cf. note 45 above). It might be noted, however, that in most cases, it is both good to have a certain basic need satisfied and bad not to have it satisfied. If we restrict our attention to these cases, we can see that several combinations are possible: (i) It may be nonderivatively bad for me not to have a certain need satisfied, but only instrumentally good to have it satisfied (e.g., "the need to form social identities"); (ii) it may be nonderivatively good for me to have a certain need satisfied, and nonderivatively bad not to have it satisfied (e.g., the need for warm human relationships); and (iii) it may be instrumentally bad for me not to have a certain need satisfied, and instrumentally good to have it satisfied (e.g. the need for a certain minimal level of income). However, I don't think that there are any needs which it would be nonderivatively good to have satisfied but "only"

Another conclusion that can be drawn is this: It can only be nonderivatively good for a person to have a basic need satisfied (*qua* need) if the following two conditions are met:

(i) The goal of the need has final value for the person. That is, if the goal of a basic need is merely instrumentally good, then the object of the need can not (*qua* object of need[51]) have final value for the needing person.

(ii) The object of the need is a constitutive part of the goal of the need, i.e. the relation between the object of need and the goal of need is a part-whole relation. That is, if the object of the need is merely a means to the goal of need, then it can never (*qua* object of instrumental need) have final value for the needing subject.

All this suggests that as far as basic needs are concerned, to know what we need is to know what is good and bad for us. However, this does not imply that we have to know what has *nonderivative* value for us in order to know what we need (in the basic sense). First, the goals of our basic needs do not always have final value for us (e.g., as in the case of survival and health). And second, there are (I think) a number of things which (like Rawls' "primary goods") can be classified as "all-purpose means", e.g., self-respect. These things are most probably objects of basic need, since they are (most likely) means to (or constitutive parts of) several goals of basic need at the same time. For these reasons, it is not really necessary to know what is prudentially valuable in order to find out what one needs. (It is also possible to have complete knowledge about what has nonderivative value for us and, due to ignorance about human nature, still not know very much about what we need).

Criticism of the basic need account

So, now that we are familiar with what a basic need is, we have to ask ourselves whether it is ever possible to justify claims about what has nonderivative value for us in terms of basic needs: Can claims of the form "situations of type X are nonderivatively good for all human beings" ever be justified by referring to the alleged fact that X is an

---

instrumentally bad not to have satisfied.

[51]Suppose, for example, that we need to be happy in order to be healthy. Happiness may well have final value, but *qua* object of basic need it is merely a means to health, i.e. something instrumentally valuable.

object of basic need? Is it ever plausible to claim that something is non-derivatively good for all of us *because* it (or something of which it is an instance) is an object of some basic need?

Well, for this to be at all possible, we have to restrict our attention to those basic needs which it is nonderivatively good for us to have satisfied, i.e. to those needs which meet conditions (i) and (ii) on p 338. Let us (for the sake of argument) assume that the need for warm human relationships is such a need, i.e. that it has nonderivative value for us to be related to other humans in this way. Is it, in this case, plausible to claim that warm human relationships are nonderivatively good for us *because* this is something we all need (in the basic sense)? I think not. Even if the object of a certain basic need happens to be nonderivatively good for a person, it cannot be concluded that it is good for him *because* he has a basic need for it. It is (rather) the other way around: in order to determine what the *relevant* basic needs are (where a need cannot be relevant unless it satisfies conditions (i) and (ii)), we must already know what has nonderivative value for us, and we can not get this knowledge by thinking about what we need. If we have the relevant notion of need in mind, we can say (somewhat figuratively) that it is not need that is prior to value; it is rather value that is prior to need. Or in somewhat different terms (I think the argument is really the same): Claims about universal prudential value can (as condition (i) shows) only be justified in terms of a certain need on the assumption that the goal of this need has final value for us. However, whether this goal is (in fact) good for us is *not* anything which can be determined by appealing to the notion of need. This point can also be expressed as follows: Suppose that it makes perfectly good sense to claim that X is good for P *because* P needs X, and that P needs X *because* X is conducive to something else (Y) that is good for P. In this case, it is the fact that Y is good for P which can not be explained in terms of needs. In particular, it seems impossible to determine whether it is nonderivatively good to have a basic need satisfied (and how good it is) without knowing how it affects the overall quality of life. That is, the notion of "intrinsic basic need" is not, as the basic need account requires it to be, "independent of our general conception of prudential value" (cf. Griffin (1986), p 52). And there is (moreover) no way in which such an account can convince us that there are certain things which are nonderivatively good or bad for us, "whether or not we would want to have the good things, or to avoid

the bad things". In short, the notion of need does not have any "justificatory force" at all, and it can therefore be eliminated altogether from this investigation[52].

The second Aristotelian appeal to Human Nature: On why filia is nonderivatively good for us

The Aristotelian argument that I have in mind is not really (at least not explicitly) an attempt to justify the idea that it has *nonderivative value* for a person to have friends. The conclusion of the argument is, rather, that the happy man (or the "supremely happy man"; the *makarios*) *needs* friends (*filoi*). On Aristotle's view, "it seems strange, when one assigns all good things to the happy man, not to assign friends, who are thought the greatest of external goods" (*NE*, IX.9, p 238). And a little later, he presents the famous argument:

> Surely it is strange /.../ to make the supremely happy man a solitary; for no one would choose the whole world on condition of being alone, since man is a political creature and one whose nature is to live with others (*NE*, XI.9, p 238).

Or, if we turn from Ross' translation to Nussbaum's:

> And surely it is peculiar to make the *makarios* a solitary; for nobody would choose to have all the good things in the world all by himself. For the human being is a political creature and naturally disposed to living-with (Nussbaum (1986), p 350).

---

[52]However, it might still be objected: "Think of the relation between rights and duties. Right-statements are often extensionally equivalent to duty-statements (e.g., it may well be the case that a certain person has a right to live if and only if it is prohibited for everybody else to kill him). However, this is perfectly consistent with the plausible idea that duties can be explained in terms of rights, viz. because rights can be conceived of as legitimate claims, i.e. as something which are attributed to the patient rather than to the agent. Now, isn't this exactly how (certain relevant) needs and prudential values are related to each other? And isn't it plausible to assume that prudential value can (for this reason) be explained in terms of need, and that the reason for this is that needs are (like rights) attributed to people in a way that prudential value is not?". This is a bad objection, however, viz. because it is based on the false assumption that to need something (in the relevant sense) is to be in a certain state. But just as the structure of prudential value is relational, so is the structure of need (with the difference that while the former relation is merely binary, the latter is 3-ary, or maybe even 4-ary; cf. note 39 on p 331).

Let us follow Nussbaum (1986) and assume that what Aristotle purports to show here is that *filia* (and membership in a political community) are intrinsically valuable for us. Can this claim be supported by appealing to our "political nature", and if so, how? Exactly how does an appeal of this type work (if it works) in this type of evaluative context?

The structure of the argument can be represented as follows:

(1) Human beings are (by nature) social beings (political creatures) who are naturally disposed to "living-with".

(2) The solitary life is not a "recognizably human life".

(3) The solitary life is a bad life.

(4) *Filia* (and/or membership in a political community[53]) is non-derivatively good for us[54].

The first question that this gives rise to is how (1) should be understood, i.e. what it means to say that "the political" is a part of human nature. For example, is it a descriptive or an evaluative (or normative) claim?

Nussbaum (1986) clearly regards (1) as evaluative, there are even reasons to believe that she regards it as equivalent with (3), or even with (4). On her view, "[t]he claim that the political is part of our nature appears to be *equivalent to* the claim that a life without it [i.e. a solitary life] is lacking in an important good, is seriously frustrated or incomplete" (p 350, my italics), and she also seems to think that this claim (i.e. (1)) "is *not a separate point from* the point about intrinsic worth or value. /.../ It is just another way of putting the point that a life without *philia* is radically lacking in essential human values" (ibid., p 367, my italics). This is how she expounds (in more detail) the idea that "the human being is a political creature and naturally disposed to living-with". This is not an appeal to

> some separate realm of natural fact, [she writes,] but to our deepest judgments of value: the solitary life is insufficient for *eudaimonia*

---

[53]In the following, I will restrict my attention to *filia* only.

[54]Where (1) is a premise (in the proper sense), where (2) and (3) are "bridging principles" (or "intermediary steps"), and where (4) is the conclusion. My reason for representing the argument in this way (rather than in terms of premises and conclusion) is that it is more "faithful" to the discussion.

> because we would not find such a life choiceworthy or sufficient for us. The solitary view of *eudaimonia* is at odds with the choices we make and the beliefs that we share. If *eudaimonia* is to include every value without which a life would be judged incomplete, it must include the political as an end in its own right. The sentence about our political *nature* indicates to us, furthermore, that political choices and concerns lie so deep that they are a part of what we *are*. The solitary life would not only be less than perfect; it would also be lacking in something so fundamental that we could hardly call it a human life at all. The appeal to nature thus underlines the depth and importance of the element in question. Without it we are not even ourselves (ibid., p 350)[55].

Now, Nussbaum may well be right about all this, but if she is, the conclusion is already embedded in the premises[56], and the argument must (for this reason) be rejected. For the argument to work, it seems that we have to understand (1) as a descriptive claim.

Now, this is exactly what Cooper (1980) seems to do. Or more specifically, he either takes (1) to mean that we are not "psychologically sufficient", or he regards it as grounded in this fact. On my view, the second interpretation is more plausible, i.e. we should regard the (natural) fact that we "suffer from" a certain kind of vulnerability or weakness as an *explanation* of the fact that we are "naturally disposed to living-with"[57]. Cooper writes:

---

[55]Now, it seems possible to give this a preferentialist (desire theoretical) interpretation: "That we are political creatures *means* that we regard the solitary life as incomplete (in a value-laden way), i.e. as *bad for us*, and that this negative evaluation is "deep", or part of our "human identity". And solitary lives are incomplete, or bad for us *because* we think so (in this deep way)". This interpretation is not accurate, however: The use of preferentialist language does not make the argument preferentialist (cf. how Nozick expresses himself when he formulates his experience machine argument), and it should also be noted that Nussbaum does not say that Aristotle appeals to *the fact that* we evaluate the solitary life in a negative way, but to the evaluation itself (whatever that means).

[56]As David Charles points out (in *The Oxford Companion to Philosophy*, on p 55): "On occasion, Aristotle seems to found his account of the good life on background assumptions about human nature, but elsewhere bases his account of human nature on what it is good for humans to achieve".

[57]This is something that Nussbaum too would agree with. At this point, it is also worth noting that the explanation in question is (most probably) evolutionary rather than straight-forwardly psychological (e.g., phenomenological), i.e. it seems fully compatible with the fact that the desire for "living-with" is (from the point of view of the desiring subject) a "given".

> If human nature were differently constituted, we might very well be immune to the uncertainties and doubts about ourselves which, according to Aristotle, make friendship such an important thing for a human being. /.../ [W]e need each other because as individuals we are not sufficient - psychologically sufficient - to sustain our own lives (p 331).
>
> To argue thus the need of human beings for friendship from the deficiencies in our psychological makeup both illuminates the nature of friendship and gives what I think is an entirely accurate account of its status in human affairs. Properly understood, there is nothing in this that should be construed as undermining or detracting from the intrinsic goodness, for human beings, of friendly relations with others. For Aristotle's point is that the deficiencies that make friendship such a necessary and valuable thing are inherent in human nature itself (ibid., p 331).

Now, all this may well be true, and it may also explain why we tend to evaluate the solitary life in such a negative way, but is it a good (objectivist) *justification* of (3), i.e. the evaluative claim that it is bad for us to live solitary lives? I think not. What Cooper says does not give any support to the idea that the *happy* solitary life is a bad life; what he shows is (at best) that such a life is impossible (or at least very unlikely).

So, if the argument does not work unless we understand (1) as a descriptive claim, and if we can not understand it as a claim about our psychological insufficiency, how should it be understood? Well, this is one possibility: (1) can be regarded as equivalent with (2), which is (in turn) understood as a descriptive claim. Now, we have already seen that

(2) The solitary life is not "recognizably human"

can be interpreted in an evaluative way, i.e. as in Nussbaum's

(2') The solitary life is "radically lacking in essential human *values*".

To see what a descriptive interpretation of (2) would look like, consider the following statements:

(i) A life without suffering is not a recognizably human life, and

(ii) An eternal life is not a recognizably human life[58].

Now, it seems rather obvious that (i) does not (at least not in itself) constitute a good reason for regarding a life free from suffering as a bad (human) life, and in a similar way, (ii) does not give us a reason to believe that an eternal life is a bad life[59]. This suggests that if statements of the form "a life without X is not recognizably human" are interpreted in this non-evaluative way[60], they can not be used as reasons for statements of the form "a life without X is bad". (And if we accept the idea that mortality is a part of human nature, we can also conclude that there are parts of human nature which are (in this context) irrelevant, i.e. that there are "human-making features" which are not "good-making features"; we will return to this point on p 349 ff.). It also allows for the possibility that it is better for a person to live a life that is not recognizably human (in this sense) than to live a life that is recognizably human.

This seems to mean that if the idea that the solitary life is not recognizably human is understood in the non-evaluative sense above, then it can not constitute a reason for accepting the idea that the solitary life is bad. Now, this strongly suggests that (3) can not be justified in terms of (2) unless it is assumed that it is good for us to have relationships with other human beings, and this is not anything that can be established by appealing to what is human and not. Or more generally: Statements of the form "X is good for all human beings" can not be justified in terms of human nature (in this "second Aristotelian way") unless it is assumed that the part of human nature to which one refers is a relevant part[61], and whether a part of human nature is relevant or not is not anything

---

[58]It is (I think) plausible to assume that both statements are true. This would mean that if our mortality is a part of our human nature (if it is a "human-making characteristic"), while our suffering is not, then (ii) is a conceptual truth, while (i) is a universal empirical truth.

[59]There are other reasons for regarding an eternal life as bad, though.

[60]This does not imply that (i) and (ii) are descriptive, however. On my view, they are (in a certain non-statistical sense) normative; a person who never suffers deviates from "the norm", i.e. he is (at least in this respect) not normal. But to say about someone that he is not normal in this sense is not necessarily to make an evaluative claim, and it allows for the possibility that non-normal people are better (or better off) than normal people.

[61]To require that it is a "good part" would be too much. In general, it seems that all talk about "good human nature" and "neutral (or bad) human nature" is (at least in this context) clearly nonsensical.

that can be determined by appealing to the notion of human nature itself (cf. note 56).

However, it is not just the step from (2) to (3) that is problematic, but also the step from (3) to (4). That the solitary life is bad does not imply that *filia* is good. What it implies is that a life can be good only on condition that it is non-solitary, i.e. if the person who lives it has relationships with other human beings. But this is, in itself, no argument for the thesis that *filia* is a necessary constituent part of a good life. However, if it assumed that *filia* is the best kind of human relationship (and that this can be shown in some way), then (4) (or something like it) follows from (3). We would then get the following inference:

(P1) The solitary life is a bad life, i.e. some form of human relationships is an essential part of the good life.

(P2) *Filia* is the best kind of human relationship.

(C) *Filia* is an essential part of the best life, and therefore good[62].

To conclude, the argument does not work: the idea that *filia* is nonderivatively good for us cannot be (objectively) justified in terms of the alleged fact that human beings are (by nature) political creatures who are naturally disposed to "living-with".

Griffin's human nature account

As we have already seen (on p 310), Griffin's (1986) (and (1996)) attempt to justify his own list of prudential values - i.e. accomplishment, active living, understanding, enjoyment, and deep personal relations - is both object-oriented and subject-oriented. Now, the subject-oriented part of Griffin's justification is clearly a kind of "human nature account". For example, he claims that all the prudential values listed by him "rest on general features of human nature; for example, autonomy is central to living a *human* existence, and we all value our humanity" (Griffin (1986), p 70), and that "some such values are clear enough features of human nature that to deny them would be a quite plain error" (Griffin (1996), p 53).

---

[62]But does this show that filia is *nonderivatively* good for us? Well, not if the notion of contributory value is intelligible *and* something can have contributory value without having nonderivative (or final) value. It is not likely that *both* these conditions are satisfied, however.

The claim that certain prudential values *are* features of human nature - and not just based on such features - makes one wonder how Griffin conceives of human nature. What is (on his view) the stuff that human nature is composed of? Well, the answer seems to be *aims* and *interests* (where prudential values are regarded as that which meet basic human interests). For example, he claims that "certain biological *aims* - for food, health, protection of our capabilities - and certain psycho-biological *aims* - for company, affection, reproduction" are "particularly deeply embedded in us" (ibid., pp 53-54, my italics), and he also claims that certain "non-biological *interests*, such as accomplishment and deep personal relations, are as firmly embedded in human nature as biological ones are" (ibid., pp 54-55, my italics). In connection with our biological nature, he also talks in terms of *needs*, e.g., when he states that some prudential judgements are "based on biological *needs*" (ibid., p 53, my italics).

So, how do we know what "interests" (or "aims") that are embedded in human nature? Let us first consider our *biological interests* (or needs), e.g., nourishment and nurture. In virtue of what are things like nourishment and nurture interests, and why should they be regarded as parts of human nature? Griffin's answer to this question is that they are human interests in virtue of "their link to the avoidance of ailment, pain, or malfunction" (ibid., p 53). Moreover, "[n]ourishment and nurture are *valuable because* they are particular forms of avoiding those core-disvalues" (ibid., p 53). (Cf. the idea (on pp 335-336 above) that we have a need for something if and only if it is necessary for the avoidance of harm).

If we turn from our biological interests to our *non-biological interests* (which is of more interest in this context, considering that all the prudential values on Griffin's list are of this type), we can see that Griffin's view is fundamentally the same. In this case too, something is an interest, and a part of human nature, in virtue of its link to the avoidance of certain kinds of harm; or more specifically, because it is a form of avoiding certain kinds of suffering, ailment, or malfunction. There are certain differences between biological and non-biological interests, however. As Griffin puts it ("roughly"), "biological ones are embedded in our animal nature and non-biological ones in our rational nature" (ibid., p 55). But more importantly, "the shift from biological to non-biological interests brings with it /.../ a shift from predominantly

experiential sorts of harm such as pain and ailment, which are fairly easily identified, to non-experiential sorts of harm" (ibid., p 55). For example, "humans are by nature sociable; they aim at love, affection, or at least company, and lack of these produces *its own sorts of pain and malfunction*" (ibid., p 52, my italics). The pain which is characteristically produced by a lack of deep personal relations is, moreover, "less experiential" than physical pain: "Without deep personal relations an adult will suffer, but it is an altogether less experiential, more contentious sense of 'suffering'[63] than the gross ailment and malfunction in the case of a baby deprived of nurturing" (ibid., p 53)[64].

As it stands, this argument gives rise to at least two sets of questions: (1) How are we to identify the "core-disvalues" that Griffin talks about? Are these disvalues are embedded in human nature too? It seems not[65], i.e. they probably have to be identified in some other way, but how? Well, it doesn't seem too hard to identify the relevant kinds of "experiential" harm (pain, ailment, and suffering); here, we can (at least in part) rely on our "hedonistic evaluations". But how are we to identify the relevant kinds of "non-experiential" harm (malfunction, ailment, and suffering[66])? And how are we to determine whether the relevant kinds of harms (experiential or non-experiential) are non-derivatively bad for us, or whether they are "merely" instrumentally bad for us? (2) How exactly are the relevant human interests linked to the avoidance of harm? And more importantly, how exactly is the positive value of having an interest met connected to the positive value of avoiding harm, and how it is connected to the negative value of the harm avoided? For example, is the positive value of having deep personal relations *derived* from the negative value of the harm which is characteristically produced by the lack of such relations? Is it good to have this interest met *because* a certain type of harm is (thereby) avoided?

These are difficult questions, but I tend to believe that at least some

---

[63]It is worth noting that in the present context, the term "suffering" is not used in the same way as it has been used so far. For example, it seems possible to suffer (in Griffin's less experiential sense) without knowing it.

[64]And "[a]s rational beings, we have an interest in accomplishment; without it life is empty in certain ways" (ibid., p 59), and so on. This suggests that all the different prudential values (i.e. their lack) correspond to different kinds of harm.

[65]Cf. the idea that it is the objects, and not the goals, of basic need that "flow from human nature".

[66]Cf. note 63 above.

of the "core-disvalues" are "merely" instrumentally bad (e.g., certain kinds of malfunction), and (more importantly) I think we can safely assume (i) that it can not be nonderivatively good for a person to have an interest met if the harm avoided is only instrumentally bad, and (ii) even if the harm avoided is nonderivatively bad, this does not imply that the "particular forms of avoiding" this harm are nonderivatively good. This suggests that Griffin's argument does not really show us that things like accomplishment and deep personal relations have *non-derivative* value. (In fact, if we ignore the bit about non-experiential harm, a hedonist may well agree with all that he says). However, on my view, the argument succeeds in justifying that deep personal relations (etc.) are good for us (but only on the assumption that the corresponding forms of harm are bad for us, and this is not anything which can be established by appealing to human nature).

To sum up, none of the four objectivist accounts which we have looked at so far has been able to meet the challenge on p 317, i.e. we have not yet come across any plausible objectivist justification of the relevant universal claims.

So, what can we learn from all this? Well, there are at least good reasons to suspect that all that a theory of human nature can show us is (at best) that certain things are good and bad for us, and that it can only do so given certain evaluative assumptions that are not themselves derived from such a theory. That is, the suspicion is that no theory of human nature can (in itself) show us what has *nonderivative* value for us, and that no such theory can be used to identify such things as "core-disvalues" or "goals of basic need".

So, is this (general) suspicion well-founded or not? Is there any reason to believe that there is some (any) human nature account that can provide an objectivist justification of the relevant universal claims? And what would such an account have to be like? In order to find out about these things, we have to conduct a more general discussion of conceptions of human nature and their possible evaluative relevance.

A general discussion about Human Nature and its evaluative relevance

So, what evaluative conclusions (if any) can we draw from a correct conception of human nature (assuming that there is such a thing)? In particular, is it possible to justify a non-internalist pluralist theory of prudential value in terms of a conception of human nature? Can claims about human nature ever function as reasons for the relevant universal claims about prudential value?

In order to find answers to these questions, we have to know what a conception (or theory) of human nature is, e.g., what kind of theory it is, and how it differs from a broader theory of man. For example, what questions does a theory of human nature purport to answer, and how does it try to answer these questions? And most importantly, what types of claims about human beings does such a theory make?

In my terminology, a theory of human nature is (roughly) a theory which purports to tell us what it is *essentially* to be human, and what it is essentially to live a human life, i.e. what it is that makes human beings (and their lives) human. Or in semantic terms (assuming that it is appropriate to treat our essential "human-making" characteristics as defining characteristics): A theory of human nature is a theory which purports to give a correct *definition* (or demarcation) of the concept "human" (or the concept of man), i.e. which purports to tell us what human characteristics that should be regarded as defining characteristics[67]. This

---

[67]It is worth pointing out that there are at least three major ways in which this central question can be interpreted: (i) The traditional ("narrow") interpretation of the question is based on two assumptions, viz. (a) that the definition of a word properly consists of expressions naming the *genus* to which something belongs and its *differentia specifica*, or distinguishing features, and (b) that the genus of man is animal (or mammal). If we combine these assumptions, we get the idea that all our human-making features are features which distinguishes us from other animals, and the question "what is it essentially to be a human?" is understood as a question about how humans differ (essentially) from other animals. (ii) The "broad" interpretation of the question is based on the assumption that a human characteristic (e.g., a need or an ability) should count as essential (or human-making) if it has "an important influence" on what it is like to live a human life (cf. Furberg (1975), p 133). This allows for the possibility that features like bodily needs and mortality are essentially human, i.e. a human feature need not (on this view) be a distinguishing feature in order to count as human-making. (iii) The third interpretation (which may be extensionally equivalent with the broad interpretation) assumes (like the traditional view) that defining features are distinguishing features, but it rejects the view that the only interesting features are

suggests that theories of human nature are metaphysical (or semantic) rather than "straight-forwardly empirical". Such theories are not totally devoid of empirical content, however, and it might also be added that they may well make evaluative claims, or claims about what is normal and abnormal (or "odd")[68].

For the sake of the argument, I will assume that there is such a thing as an objectively true conception of human nature, and that it is (in principle) possible to determine what our human nature consists in. It is worth noting that these are problematic assumptions, however, e.g., for the following two reasons:

(i) Any theorist of human nature must (so to speak) make up his mind about the following issue (formulated by Michael Ruse in *The Oxford Companion to Philosophy*, p 376):

> Must human nature be defined with respect to the new-born infant, in which case it would seem to be a bundle of potentialities, or is it to be defined with respect to the full-grown adult, in which case does one consider training something crucial to the development of human nature or is it rather something which takes our nature from its true state?

Now, the mere fact that such a decision has to be made suggests that there are no such thing as *the* objectively true conception of human nature[69]. And if we, on top of this, assume that human nature is "to be defined with respect to the full-grown adult" (this is an assumption that

---

the features which distinguishes us *from other animals*. On this view, the concept of man should not just be distinguished (demarcated) from concepts like "ape" and "dog", but also from concepts like "angel", "ghost", "advanced computer", "android", and "god". This can also be conceived of as an *expansion* of the traditional *genus* of man: the *genus* of man is no longer animal (or mammal), but *real or fictional "living" creature*. That is, on this view, the question is not just a question of how we differ from other animals, but also how we differ from angels, gods, and the like, and this suggests that the number of human-making features may be pretty large. (It is worth noting that as it has been described here, this is not a very fruitful approach. For example, it can hardly be assumed that every single difference between humans and some other creature should be regarded as human-making. So, how do we determine what differences that are of significance? Well, here it seems natural to turn to (ii)). In short, it is highly likely that (ii) is (in this context) the best interpretation.

[68]I suspect that theories about masculinity and femininity share at least some of these features with theories of human nature, especially the last one.

[69]Isn't there something odd about trying to find the essence of something which changes so much over time as a human being?

is almost always made in this particular context), it can be argued that it is even less likely that there is such a thing as *the* correct conception of human nature; especially if training (which is a cultural thing) is regarded as "crucial to the development of human nature"[70].

(ii) On the assumption that it is at all possible to determine what our human nature consists in: How do we do this? Or alternatively, how are we to determine whether a certain theory of human nature is good or not, and whether one of two competing theories of human nature is better than the other? Well, it is possible that some of the claims made by such theories can be verified or falsified by "observation" (in a wide sense of the term), e.g., by facts about how people are (and were) living, what people actually like and value, and so on. But it is also clear that there are many claims about human nature which can neither be verified nor falsified by observation. (For example, the fact that hermits exist does not falsify the claim that human beings are social beings, and the fact that some people are utterly irrational does not falsify the claim that human beings are rational beings). But if we cannot appeal to observation, and if we are (in this context of justification) required to remain value-neutral[71], what can we do?

Let us now turn to the feature of "human nature accounts" which is (in this context) most important, viz. their content. The question is: What kinds of statements are typically included in a theory of human nature? What kinds of different claims about human nature are there (apart from the claims with which we are already familiar, e.g., claims about universal human needs, or claims about what the ergon of man consists in)? And above all, do any of these claims have any "justificatory power"?

Here are some examples of types of claims which are (intuitively)

---

[70]At this point, some "anti-essentialists" ("constructivists", or "historicists") would probably appeal to the mere fact that there is an enormous cultural and historical diversity. This fact does not constitute any serious threat to the idea of a common human nature, however (e.g., to the idea that there are a number of basic biological and non-biological needs); at least as long as we allow enough room for human plasticity. It can also be argued that the training given in some cultures is better than the training given in other cultures, viz. because it is more conducive to the development of human nature! (cf. note 124 on p 274 above).

[71]Now, as we will soon see, this may well be too strong a requirement: Perhaps the important thing is not that we are value-neutral, but that we are value-neutral in the right way (or ways), e.g., that we do not appeal to intuitions about what is good and bad for us.

about human nature:

(1) Claims about what is innate, as opposed to what is "culturally acquired", e.g., "human nature is inherently good", "human beings are by nature aggressive and disposed to conflict (or curious)".

(2) Claims about what is normal and abnormal (and not just in a statistical sense), e.g., "homosexuality is against human nature"[72].

(3) Certain types of claims about what is distinctively human, e.g., "Homo sapiens is the only species which is capable of play (art, religion, science, war, torture, stamp-collecting, or ice-hockey)".

(4) Different kinds of claims about human motivation, e.g., "ultimately, we do what we do in order to achieve pleasure and avoid suffering" (psychological hedonism) or "human beings are lazy (or curious) by nature".

(5) Claims about how we are "shaped" or "formed", e.g., "human beings are, to a considerable extent, victims of circumstance and coincidence, i.e. they are shaped by circumstances which are (were) not under their control"; and claims about how "shapable" we are, e.g., "human nature is characterized by a high degree of plasticity" (an idea that is typically invoked in order to explain why human life can take so many different forms).

Now, it is clear that none of these claims can be used to justify (in an objectivist way) the relevant universal claims about prudential value. For example, the alleged fact that there is something odd about a person whose sexual desires are directed towards members of his own sex does not give us a reason for believing that it is bad for particular homosexuals to engage in homosexual behaviour (and neither does the alleged "fact" that homosexuality is, in the individual case, wrong). The only claim above which have some justificatory power is psychological hedonism, and the reason for this is that it contains the idea that we all have a *desire* to feel pleasure and an *aversion* to pain and suffering. But this is a subjectivist idea, and it is (as such) of no help to the objectivist.

However, it should not surprise us that none of these claims have any justificatory power. After all, the theories of human nature of which these claims are constituent parts are all (with the exception of psychological hedonism) constructed for "non-evaluative purposes". This

---

[72]This particular claim can also be interpreted in other ways, e.g., as a normative claim, i.e. as "homosexuality is wrong".

suggests that we should instead focus our attention on theories of human nature which are explicitly constructed for evaluative (or other ethical) purposes. A good example of such a theory (a theory which is, so to speak, broader than the theories we have already looked at, and therefore "more illustrative") is Nussbaum's (1990) "thick vague conception of the human being"[73].

According to this "Aristotelian" theory about "what it is to be situated in the world as human", there are 11 features which should be regarded as "part of any life that we count as a human life" and "constitutive of [our] humanness", viz. the following ones:

(1) *Mortality*. "All human beings face death and, after a certain age, know they face it. This fact shapes more or less every element of human life. Moreover, all human beings have an aversion to death" (p 219).

(2) *The Human Body*. "We live our entire lives in bodies of a certain sort, whose possibilities and vulnerabilities do not as such belong to one human society rather than another", and these bodies constitute a "deep demarcation" of our possibilities. The following human-making features are, on Nussbaum's view, the most important aspects of our "bodiliness": (a) "Hunger and thirst; the need for food and drink"; all human beings need food and drink in order to live, and "all human beings have appetites that are indices of need"; (b) "Need for shelter"; which is based on our "relative fragility and susceptibility to heat, cold, the elements in general"; (c) "Sexual desire"; and (d) "Mobility"; human life is "in part constituted by the ability to move from place to place in a certain way", and "human beings like moving about, and dislike being deprived of mobility" (ibid., pp 220-221).

(3) *Capacity for Pleasure and Pain*. "Experiences of pain and pleasure

---

[73]It is important to note that this theory is not primarily constructed for the purpose of justifying a theory of prudential value, but for "political purposes". On her political view, it is the task of the government (the legislator) to promote the good of human beings, and the kind of good that should primarily be promoted is "basic human functional capabilities" (cf. p 225). (That is, "it is capabilities, not actual functionings, that should be in the legislator's goal" (ibid., p 224)). This is where the theory of human nature (or "Level A" of "the thick vague conception") enters the picture; it is used to specify a list of basic functionings "that should, as constitutive of human life, concern us" (this is the second stage of "the thick vague conception"), which is then (in turn) used to "derive" a list of 11 basic human functional capabilities (which Nussbaum regards as "a minimal theory of good" (ibid., p 225)).

are common to all human life. /.../ Moreover, the aversion to pain as a fundamental evil is a primitive and, apparently, unlearned part of being a human animal" (ibid., p 221).

(4) *Cognitive Capability: Perceiving, Imagining, Thinking.* "All human beings have sense-perception, the ability to imagine, and the ability to think, making distinctions and, as Aristotle famously says, 'reach(ing) out for understanding'. And these abilities are regarded as valuable" (ibid., p 221).

(5) *Early Infant Development.* "All human beings begin as hungry babies, aware of their own helplessness, experiencing their alternating closeness to and distance from that, and those, on which they depend. This common structure to early life /.../ gives rise to a great deal of /.../ experience that is of great importance for the formation of emotions and desires" (ibid., p 221).

(6) *Practical Reason.* "All human beings participate (or try to) in the planning and managing of their own lives, asking and answering questions about what is good and how one should live. Moreover, they wish to enact their thought in their lives - to be able to choose and evaluate, and to function accordingly" (ibid., p 222).

(7) *Affiliation with Other Human Beings.* "[A]ll human beings recognize and feel a sense of affiliation and concern for other human beings. /.../ Moreover, we value the form of life that is constituted by these recognitions and affiliations - we live to [sic] and with others, and regard a life not lived in affiliation with others to be a life not worth living" (ibid., p 222).

(8) *Relatedness to Other Species and to Nature.* "Human beings recognize /.../ that they are animals living alongside other animals, and also alongside plants, in a universe that, as a complex interlocking order, both supports and limits them. We are dependent upon that order in countless ways; and we also sense that we owe that order some respect and concern" (ibid., p 222).

(9) *Humour and Play.* We are "the animals who laugh", "[i]nability to play or laugh is taken, correctly, as a sign of deep disturbance in an individual child", and an "entire society that lacked this ability would seem to us both terribly strange and terribly frightening". Nussbaum concludes by claiming that we "certainly do not want a life that leaves this element out" (ibid., p 223).

(10) *Separateness.* "However much we live for and to others, we are,

each of us, 'one in number', proceeding on a separate path through the world from birth to death. Each person feels only his or her own pain and not anyone else's. Each person dies alone". And if "fusion is made the goal [of human interaction], the result is bound to be bitter disappointment" (ibid., p 223).

(11) *Strong Separateness*. "Because of separateness, each human life has, so to speak, its own peculiar context and surroundings - objects, places, a history, particular friendships, locations, sexual ties - that are not the same as those of anyone else, and in terms of which the person to some extent identifies herself. /.../ [T]here is no life /.../ that really does fail to say the words 'mine' and 'not-mine' in some idiosyncratic and non-shared way" (ibid., pp 223-224)[74].

After having presented this list, Nussbaum points out that it is both open-ended and evaluative. The reason why it is *open-ended* is that "[o]ne could subtract some items and/or add others", and it is *evaluative* because "it singles out some items, rather than others, as the most important items, the ones in terms of which we identify ourselves" (ibid., p 224). But notice that this does not mean that the claims listed are, in themselves, evaluative. For example, the claim that "all human beings participate (or try to) in the planning and managing of their own lives" is not, in itself, an evaluative claim; it is (rather) the claim that this is a human-making fact which is (in a certain, presumably harmless, sense) evaluative.

So, what *types* of claims does Nussbaum's theory contain? According to Nussbaum herself, "the list is composed of two different sorts of items: *limits and capabilities*" (ibid., p 224, my italics), i.e. "limits against which we press and powers through which we aspire" (ibid., p 219). This idea is only partly correct, however. It is true that the list is, *in part*, composed of these types of items, but there are also a number of items which can not be subsumed under these categories. But before we look at these other items, let us first look at those items that can, in fact, be classified as limits or capabilities.

(i) The only claims that can be classified as claims about *limits* are (1) "mortality", (2) "the human body", (8) "relatedness to other species

[74]Much of what Nussbaum says can (of course) be questioned, but it is not really necessary for my purposes.

and to nature", (10) "separateness", and perhaps (11) "strong separateness". Some of these limits (e.g., the limits that "flow" from our bodiliness) are needs, but there are also other types of individual limits on the list, e.g., our mortality, our vulnerability, and our separateness. The fact that we are dependent on the natural order in a certain way is an "*external* limit" and should (I think) therefore be regarded a part of the human condition rather than a part of human nature (which may only include *internal* limits[75]). (Another reason for excluding our dependency on the natural order from the list is that it is a *collective* rather than an individual limit).

(ii) The only claims which can possibly be categorized as claims about *capabilities* (or abilities) are (2:d) "mobility", (3) "capacity for pleasure and pain", (4) "cognitive capability", (6) "practical reason", and (9) "humour and play". On Nussbaum's view, human beings have essentially the ability to move about, to feel pain, to perceive things, to imagine things, to make distinctions, to manage their own lives, to laugh and play, and so on. But if we take a closer look at these items, we can see that some of them include more than just capabilities. For example, it is misleading to subsume facts like "there is no painless human life" and "human beings have an aversion to pain" under the heading "the *capacity* for pleasure and pain", and item (9) is (it seems) really a claim about what is normal and abnormal.

But as has been indicated, apart from the claims about limits and capabilities, the list also contains other types of items. Here are some examples of such items:

(iii) *Claims about desires and aversions*. Nussbaum seems to regard it as essentially human to have an *aversion* to things like death, not being able to move about, pain, and a life without humour and play, and to *desire* (or like) things like food and drink, sex, moving about, and participation in the planning and managing of one's own life.

(iv) Closely related to this type of claims are the *claims about what we value*. We regard our cognitive capabilities as valuable, she says, and we regard a solitary life as a life "not worth living".

(v) *Claims about what is normal and abnormal (or "disturbed")*. A child (person) who is unable to laugh and play is considered abnormal, and

---

[75]Our *awareness* of external limits can (to the extent that this awareness is itself a limit) be regarded as internal, however.

so is (it seems) someone who does not "feel a sense of affiliation and concern for other human beings". (This latter item could also be subsumed under the category "claims about feelings").

(vi) *Claims about our individual history*, about how we have (as individuals) become what we are. We all begin as helpless babies, she says, and this is something which is of "importance for the formation of [our] emotions and desires".

Now that we have seen what types of claims that are included in Nussbaum's theory of human nature ("level A of the thick vague conception"), the question arises: Can any claims of any of these types ever be used to justify any claims about what is good and bad for us? In order to find an answer to this question, we will go through the different types of claims, one by one, to see whether any of them has any justificatory power.

(i) *Claims about limits* (e.g., claims about our needs, our mortality, our vulnerability, or our separateness) have little or no justificatory power, especially not in this context. For example, the idea that it is bad for us to die (before a certain age) can neither be justified in terms of the fact that we will in fact die (which is a limit), nor in terms of our awareness of this fact; and the idea that it is good for us to live our lives in an autonomous way does not get any support whatsoever from the fact that we cannot fuse with other people (a limit), or from the fact that we die alone. In short, it seems that claims about what is impossible for us can never function as justifications for claims about what is good or bad for us[76].

(ii) So, what about *claims about capabilities* ("human-making" or not); can such claims ever justify claims about what is good or bad for us? I think not. For example, the hedonistic idea that pleasure is good and pain bad for us can most certainly not be justified in terms of our capacity for pleasure and pain; and the alleged fact that the ability to think is essentially human does not (in itself) give any support to the idea that it

[76]The only reason for regarding knowledge about limits as relevant at all is that we have to take our limits into account when we live our lives (e.g., when we set our goals, and when we try to realize these goals), and sometimes we are also (as in the case of death) forced to "relate" to them in some way or other. (What this means is that claims of this type do not just belong to conceptions of human nature, but also to conceptions of *the human condition*, which are, to a considerable extent, conceptions about what is possible and impossible (with some emphasis on the latter)).

is good for us to think. More generally, claims about capabilities can only be used to justify claims about prudential value on the assumption that it may (at least under certain conditions) have nonderivative value for a person to exercise and/or develop his capabilities. But this assumption is (in most cases) a rather dubious assumption, and if the capabilities we have in mind are of the universal and "human-making" type, it is even more implausible.

(iii)-(iv) Claims about our *desires and aversions*, and claims about *what we value*, can (it seems) be used to justify claims about what is good and bad for us. For example, it seems that the fact (if it is a fact) that it is essentially human to have an aversion to premature death, or to pain, is not a bad reason for regarding premature death, or pain, as bad for us. But if the idea is that it is nonderivatively bad for us to feel pain, then it must (I think) be based on the claim that it is essentially human to have an *intrinsic* aversion to pain, and this idea is not as intuitively plausible as the idea that the aversion to pain (intrinsic or not) is essentially human. In any case, justifications of this type are all subjectivist[77], and they can therefore be ignored in this "objectivist context".

(v) So, can *claims about what is normal and abnormal (or "disturbed")* be used to justify claims about what is good and bad for us? I think not[78]. For example, the idea that a good life contains some laughing, playing, and recreational activities can not be justified by appealing to the idea that a child who is unable to laugh and play is deeply disturbed. First, that something counts as a disturbance in a child does not imply that it should count as a disturbance in an adult. And second, even if a certain abnormal activity (or character trait) is bad for us, this does not imply that it is the fact that it is abnormal that makes it bad. Instead, it may well be the other way around, i.e. it may well be the case that certain activities or traits that have been classified as abnormal *because* they are bad for the agent or for other people. (Cf. also the discussion of (i) and (ii) on p 344).

---

[77]But it should be noted that these justifications differ from the subjectivist justifications on pp 314-315 above: There, a reference is made to what we all (in fact) desire, but here, a reference is (instead) made to what it is essentially human to desire.

[78]At least not in this particular context. However, in chapter 3, I suggested that intuitions about what is normal and abnormal might be of relevance in a hedonistic context (cf. (ii) on p 146). So the question arises: Are these two suggestions really compatible?

(vi) *Claims about our individual history*, about how we have become what we are, can (of course) not be used to justify claims about what is good and bad for adults. For example, the idea that it is good for us (*qua* adults) to have attachments to things and persons outside ourselves, e.g., to (be able to) love those who love and care for us, can (obviously) not be justified in terms of what it was like to be an infant. (For some reason, it seems that no statement about what is good for adult human beings can be justified in this way)[79].

To conclude, it does not seem possible to provide an objectivist justification of the relevant evaluative claims in terms of human nature. With the possible exception of claims about desires and aversions (which belongs in a subjectivist context), there are (it seems) no "factual" claims about human nature which have the requisite justificatory power. Moreover, even if it were possible to justify *some* of the relevant universal claims in terms of human nature, it would hardly be possible to justify a *complete* theory of prudential value in this way. Examples of plausible non-internalist claims about prudential value which can obviously not be derived from claims about human nature are the idea that it is good for us to live autonomously, and the idea that it is (nonderivatively) good for us to be in contact with reality[80].

---

[79]But what it was like to be a child may well be something that a person should (*qua* facticity) take into account when he is leading his life. Moreover, developmental claims may also (on the assumption that it is good to be a normal adult) be used to justify claims about what is (instrumentally) good for infants and other children.

[80]Suppose that there are a number of plausible claims about prudential value (e.g., that autonomous living and contact with reality are good for us) which do not get any support whatsoever from the correct theory of human nature. If this is correct, there are things that are good for us, but not in virtue of our common nature. However, this is not inconsistent with the idea that if something is good for us, then it must (in part) be good for us because of what we are: After all, we are more than just nature. This may open up for some cultural relativism, however, viz. in the following way: "If something is good for us, then it must, to some extent, be good for us in virtue of what we are. This means that if something is not good for us in virtue of our common nature, it must be good for us in virtue of the 'cultural element' within us. And since we live in different cultures, this suggests that different things are good and bad for us". Now, suppose that this idea is correct. Does this imply that some claims about prudential value can (and must) be *justified* in cultural terms? What would such a justification look like? For example, can the claim that autonomous living is good for Westerners be justified by referring to the fact that autonomy is of great importance in the Western Culture? Well, perhaps it can, but only on the assumption that these values are so deeply embedded in us that they can be

But does this allow us to conclude that the correct theory of human nature (assuming that there is such a thing) is totally lacking in evaluative relevance? I think not. More specifically, I think there is at least one way in which a correct conception of human nature can be of relevance in this context, viz. by providing a set of minimal requirements that every plausible theory of the good life must meet. As Kekes (1988) suggests,

> [m]oderate naturalism [i.e. Kekes' own theory of human nature] influences theories of good lives by setting for them a minimum standard of adequacy; any adequate theory must do justice to the relevant physiological, psychological, and social facts. The reason for this is that since theories of good lives are concerned with human welfare, and since these facts establish the minimum requirements for the welfare of any human being, theories of good lives either take these facts into account or they cannot be adequate (pp 40-41).

But what is meant by saying that an adequate theory of the good life must "do justice to" the correct theory of human nature, or that it must take it "into account", "incorporate" it, or "recognize" it? What exactly is it for a theory of prudential value to do justice to the correct theory of human nature? Consider the following example: Assume that it is essentially human to be an agent, and to be a social creature. In this case, a theory of prudential value which claims that it is positively bad for us to be active, or to have friends, clearly fails to do justice to the correct theory of human nature.

Let us now consider the hedonistic theory. Like the desire theory, this theory allows for the possibility that totally passive lives or solitary lives are good lives. Does this mean that hedonism fails to do justice to our rudimentary theory of human nature? Well, it is clear that hedonism *as such* does not "do justice to", or take it "into account", that we are agents and social beings. In fact, it is not likely that a hedonist (*qua* hedonist) takes human nature into account at all. The only part of human nature that a hedonist has to take into account is our capacity for pleasure and pain, and he may also consider the fact that we all have (for the most part) an aversion to pain and a desire to feel plea-

---

correctly conceived of as a kind of "second nature", or part of our "identity". But even if this is so, it is doubtful whether it is really possible to justify the relevant evaluative claims in terms of such "deep cultural facts".

sure. (However, it is very likely that "in practice", a hedonist will also take our rudimentary theory into account, viz. because he wants to be informed about what it is that tends to make us happy or unhappy). But is this something that can be held against hedonism (as Rachels (1986) seems to think)? I think not. Consider the following disagreement: The pluralist claims that friendship has nonderivative value for us, while the hedonist claims that it is "merely" instrumentally good. Now, it is hardly possible to settle this disagreement by appealing to the fact that it is essentially human to have friends, and this suggests that the hedonistic theory is not really inconsistent with the idea that human beings are (by nature) social beings.

Here is another example. Assume that human beings are more than "body-minds", that we are also *spiritual beings*. (I don't really know what our spirituality is supposed to consist in, but I assume that it makes sense to talk about spiritual needs, a spiritual potential, spiritual development, spiritual awakening, and so forth). Does this constitute a threat to all "secular" theories of prudential value, i.e. theories which do not acknowledge any "spiritual prudential values"[81], i.e. things that are good and bad for us in virtue of our spirituality, or *qua* spiritual beings? Does the fact that all secular theories of prudential value[82] allow for the possibility that spiritually poor lives are good lives imply that these theories fail to do justice to the fact that we are essentially spiritual, and that they should (for this reason) be rejected? I think not. If we are essentially spiritual, then this is something that we must somehow take into account, but there are (as we have seen) several ways in which this can be done. For example, "spiritually good living" need not be regarded as something nonderivatively valuable (or as a necessary part of the good life); it may also be regarded as instrumentally valuable (or as a necessary *condition* for good lives)[83].

---

[81]I hesitate to call these values prudential, though; the term "prudential" somehow belongs to the framework of prudence, and prudence *may* be incompatible with "true spirituality".

[82]This includes all the substantive theories of prudential value with which we are familiar.

[83]As an illustration of the extent to which most theories of prudential value is based on "secular" assumptions, it can be enlightening to consider the sheer possibility that a life can be successful and pleasant, and full of creative activity and intimate relationships, and still (because of its spiritual poverty) bad, or that a life full of suffering is (because of its spiritual richness) good. (Cf. Matt. 16.26: "For what is a man profited, if he shall gain the whole world, and lose his own soul?").

To conclude, it seems that it is possible for a theory of prudential value to fail to do justice to the correct theory of human nature, but the requirement that a theory of prudential value should do justice to human nature cannot (if it is plausibly interpreted) be used to exclude any of the traditional theories of prudential value.

This concludes the general discussion of conceptions of human nature and their possible evaluative relevance. To sum up, it seems highly unlikely that there is any human nature account that can provide an objectivist justification of the relevant non-internalist claims. And since the subjectivist and quasi-subjectivist attempts to justify these claims also failed, as did the "non-subject-oriented" arguments, we can conclude that we have little or no reason to accept any of the relevant (universalist and non-internalist) claims. And since the counter-arguments against the non-internalist pluralist theories are stronger (strong enough to place the burden of proof on the pluralists), we should (I think) reject all such theories[84].

Weaker versions of "the objective list theory"

At this point, the non-internalist pluralist might concede that it is not very plausible to claim that it is good for all of us to have intimate relationships or to live our lives autonomously, regardless of whether we regard these things favourably or unfavourably. Or alternatively, he might reject the "tough-minded" (literal, or strong) interpretation of the idea that there are objective (and universal) prudential values; and he might reject the corresponding strong (pure, simple, extreme, or raw) version of "the objective list theory", according to which "being well-off is simply a matter of one's having the various objective goods" (cf. Kagan (1992), p 170), and according to which "we can measure changes in a person's well-being just by the amount that he realizes objective values" (cf. Griffin (1986), p 54)[85].

But he might still insist that there is a certain grain of truth in the

---

[84]It is worth noting that this conclusion depends entirely on the assumption that (1) is a good argument against all theories of this type. But what if this assumption is not valid? Well, this would suggest that there are no good arguments neither for nor against the "theory".

[85]Cf. p 286 above, including note 2. It is also worth pointing out that the Nozick passage quoted on p 305 is a good example of this type of view.

"theory", or more specifically, that there are certain intuitions which gave rise the different non-internalist theories in the first place, and that it might be possible to save these intuitions by introducing weaker versions of these theories; or alternatively, by making "weaker interpretations" of the idea that there are objective prudential values, and of the idea that friendship and autonomy are such values. As I see it, there are at least three ways in which this can be done, viz. the following ones:

The first two alternative interpretations of "the objective list theory" both regard the theory as a theory about "contributory value" rather than as a theory of "intrinsic value". The idea is that the presence of an "objective prudential value" (e.g., friendship) makes certain *wholes* more "intrinsically valuable" than they would otherwise have been. Depending on what kind of whole one has in mind, there are two versions of this idea:

(1) The first contributory view is based on the idea that an "objective list theory" cannot be plausible unless it is, in some way, combined with hedonism, the desire theory, or both. On this type of view, there are two ways in which the claim that it is "objectively prudentially valuable" for us to have friends can be understood:

(a) If friendship is "objectively prudentially good" and solitude is not, then it is nonderivatively better for a person to take pleasure in having friends than to take pleasure in living in isolation. That is, having friends may be contributorily valuable by making the whole "P has friends and takes pleasure in this" nonderivatively better than "P has no friends and takes pleasure in this".

(b) If friendship is "objectively prudentially good" and solitude is not, then it is nonderivatively better for a person to have his desire to have friends satisfied than to have his desire for isolation satisfied. That is, having friends may also be contributorily valuable by making the whole "P intrinsically wants to have friends and has friends" nonderivatively better than "P intrinsically wants to live in isolation and lives in isolation".

To give "objective prudential values" this kind of role is to regard all "objective list theories" as "servants" to ("modifiers" of, or "complements" to) hedonism or the desire theory. That is, an "objective list" is (on this view) nothing but a set of criteria according to which we can determine what pleasures (and what desire-fulfilments) that are

best for us. This suggests that "the objective list theory" can be regarded both as a theory about what is objectively pleasant (unpleasant), or what is worth taking pleasure (or displeasure) in, and as a theory about what it is worth desiring (or avoiding) "in the prudential sense". That is, all objective list theories can be formulated both in a hedonistic idiom and in a desire theoretical idiom. But this should not come as a surprise to us, since it has (in fact) already been done. (Cf. for example (R2) on p 145, and (iv) and (vii) on p 272).

It is also worth mentioning that if the different non-internalist pluralist theories are interpreted in this (weak) way, they do not get hit by the argument which refuted all the pure versions of the "theory", viz. (1) on p 304. Moreover, the weak versions of the relevant non-internalist theories are also highly consistent with the intuitions on which the atomist arguments on pp 309-310 are based[86], e.g., with the idea that it is (nonderivatively) better for a person to do X and choose it than to do X without having a choice, or with the idea that it is better for a person to have his desire for a love affair fulfilled than to have his desire for degradation fulfilled.

(2) On the second type of contributory view, "the objective list theory" is regarded as a theory about what contributes to the final value of a life-at-a-certain-time (considered as a whole). Here, the claim that it has "objective prudential value" for us to have friends is either understood as (a) "Whatever a person's attitudes towards friendship are; his life is better if he has friends than if he does not have any friends" (where friendship is not regarded as a necessary component in a good life), or as (b) "A life with no friendship in it cannot be a good life; friendship is a necessary constituent part of a good life".

It is hard to tell exactly how strong these interpretations are. For example, (b) (which seems stronger than (a)) is clearly consistent with (1:a) above, i.e. with modified hedonism (Cf. Nozick's and Mill's arguments on pp 137-141). As I see it, this is one good reason for ignoring interpretations of this type. Another reason is that it is not clear whether the "theory" can (if interpreted in this way) meet objection (1) on p 304.

(3) It is also possible to regard "the objective list theory" as a (partial)

[86]In fact, these "intuitions" are not just consistent with the relevant "weak objectivist" claims; they *are* themselves examples of such claims.

conception of what a good, well-functioning, human being is like. Here, the idea that it is "objectively prudentially good" for us to have friends is taken to mean that it is good for a mature, well-functioning human being to have friends. But why is this? Well, a well-functioning person (in Aristotle's sense) is a person who recognizes what is objectively valuable, e.g., he is the kind of person who recognizes that friendship is a good thing, who wants to have friends, and who takes pleasure in having friends and in spending time with them. This is how an objective list theory might be viewed as a (partial) theory of what a well-functioning person is like.

On this interpretation, the theory does not really purport to tell us what is good for all human beings: the idea that it is good for us to have friends is (on this interpretation) compatible with the idea that it may not be good for everyone to have friends (i.e. it does not get hit by objection (1) on p 304). For example, it may not be good for a hermit (who does not want to have friends, and who would not take pleasure in it) to have friends. But since it would be good for all of us to become more well-functioning, we could say that even if it would not be good for the hermit (as he is now) to have friends, it would be good for him to become the type of person who would want to have friends (etc.)[87].

To conclude: There are four possible interpretations of "the objective list theory". If the "theory" is interpreted in the first (tough-minded) way, it is implausible, and if it is interpreted in the third way (i.e. as part of a conception of the good life, considered as a whole), it is too vague. The two interpretations of the "theory" on which it is most plausible are the second and the fourth. The latter interpretation (where the "theory" is regarded as part of a conception of "the good man") is of little relevance in this context, however. This suggests that we should (in this context) regard "the objective list theory" as a theory of what is objectively pleasant (unpleasant), and/or as a theory about what is worth desiring (or avoiding) "in the prudential sense". So, is it plausible to assume that some (any) version of the "theory" (understood in this

[87]To regard the objective list theory in this way (as a part of a theory of what a well-functioning person is) is not unreasonable. In fact, this interpretation can be backed up by the fact that Aristotle's ethical theory is both an objective list theory about "the good life for man" and an objectivist theory about "the good man", and that he does not really distinguish between "the good life for man" and "the life of the good man". The reason why he failed to make the distinction may (after all) be that there isn't really much of a distinction to make.

sense) is a plausible theory of prudential value? Well, I think some version of the "objective list theory" (interpreted in this weak way) is an essential *part* of such a theory, but I am not able to offer any new arguments for this view.

This concludes the chapters on non-internalist pluralism, and it is now time to take a look at my own mixed theory.

Chapter Eight

My own mixed theory

It has already been implied that my own theory is a kind of "combined theory", or more specifically, a kind of fusion of the three theories. Now, the idea that the most plausible theory is a combined theory is not new. For example, both Sumner's (1996) and Nordenfelt's (1991) subjectivist theories can be seen as combinations of hedonism and the desire theory, and Parfit (1984) suggests that the most plausible theory might be a combination of all three theories. He writes:

> What is good for someone is neither just what Hedonist's claim, nor just what is claimed by Objective List Theorists. We might believe that if we had *either* of these, *without the other*, what we had would have little or no value. We might claim, for example, that what is good or bad for someone is to have knowledge, to be engaged in rational activity, to experience mutual love, and to be aware of beauty, while strongly wanting just these things. On this view, each side in this disagreement saw only half of the truth. Each put forward as sufficient something that was only necessary. Pleasure with many other kinds of object has no value. And, if they are entirely devoid of pleasure, there is no value in knowledge, rational activity, love, or the awareness of beauty. What is of value, or is good for someone, is to have both; to be engaged in these activities, and to be strongly wanting to be so engaged (p 502)[1].

My own combined theory is (of course) constructed in such a way so as to be able to stand up to all the objections which have been directed against the other theories, i.e. it is the only theory which is fully consistent with the intuitions on which these objections are based. But

[1]This passage has already been quoted on pp 152-153.

before I present the theory, let us first recapitulate the most important points that have been made in the earlier chapters (these points will, after all, constitute the "building blocks" of the theory). In order to simplify things, I will restrict my attention solely to the positive values.

A short recapitulation of the earlier chapters

In the chapters on hedonism, a number of anti-hedonistic claims were made: First, pleasure is not all that matters; there are other things besides pleasure (and experience) which matter to us. Or more specifically, how well off a person[2] is (on the whole) is not just dependent on how pleasant his total mental state is, but also on other things, e.g., how much desire-fulfilment there is in his life, and to what extent he lives his life autonomously. Moreover (and this is a stronger claim), it seems that certain situations have nonderivative value for a person even though they do not have any pleasant experiential content at all. And second, there are certain pleasures which are not good for us to have, e.g., pleasant emotions the intentional objects of which are "objectively unpleasant".

On the positive side, there is (obviously) some truth in the hedonistic theory. It is not just that it is almost always nonderivatively good for us to feel pleasure; pleasure is also an *important* good (for example, a life devoid of pleasure can hardly be regarded as a good life). Moreover, it seems plausible to assume that a person's well-being can not be directly affected (at least not for the better; cf. pp 268-269 above) by things he does not know anything about.

I have also suggested that the most plausible version of the hedonistic theory is a modified version which includes certain "objectivist elements", viz. the following ones:

> (R2) If the object of a pleasant emotion is an "objectively unpleasant" situation, e.g., a situation that is not worth taking pleasure in "in the prudential sense" (e.g., being humiliated, deceived, or dominated); then it is not nonderivatively good for the subject to have the

[2]An adult and normal human being, that is. If we are interested in how to determine how well off an infant is, or a mentally retarded human being, or a squirrel, the hedonistic theory may be somewhat more plausible. This suggests that it is highly problematic to make interpersonal comparisons of well-being between normal adults and infants (etc.).

emotion.

(RW3) If the intentional object of a pleasant emotion E1 is objectively valuable (e.g., in the prudential sense), while the object of another pleasant emotion E2 is not, then it is *ceteris paribus* nonderivatively better for the subject to have E1 than to have E2[3].

I have also claimed (RW2) that it is *ceteris paribus* better to have pleasant emotions that are based on true beliefs than to have pleasant emotions that are based on false beliefs. That is, the complex situation (conjunction)

P takes pleasure in X[4]
P believes that X[5]
X obtains

is better for P than the conjunction

P takes pleasure in X
P believes that X
X does not obtain

In the chapters on the desire theory, I suggested that it is often, but not always, good for us to have our intrinsic desires fulfilled. Or more specifically, it is nonderivatively good for a person to have an intrinsic desire fulfilled if and only if the desire is relevant (it is about the subject's own life, it is rational in the relevant sense, and the subject is aware of having it) *and* the subject is aware of the fact that the object obtains (the desired situation "enters" the subject's experience)[6]. That is, it is not just (in the case of relevant desire) that

---

[3]To give "objective prudential values" this kind of role is (as we have seen on pp 363-364) to regard "the objective list theory" as a theory about what is worth taking pleasure (or displeasure) in ("in the prudential sense"), i.e. as a "modifier" of, or "complement" to hedonism.

[4]Where the fact that P takes pleasure in X does *not* imply that X obtains, i.e. where X is (merely) an intentional object of the pleasant emotion.

[5]Where this "belief" need not be propositional. The reason why I have not used the phrase "P is aware of X" is that it suggests that X obtains (viz. because we tend to regard "being aware of" as a so-called success verb).

[6]However, it *may* be bad for a person to have his aversions fulfilled, even though he is not aware of the occurrence of the object (cf. pp 268-269 above).

P desires that X X obtains P believes that X

is better for P than

P desires that X X obtains P does not believe that X

situations of the latter type do not have any positive nonderivative value for P at all[7].

Moreover, desire-fulfilment is not the only thing that is good for us; it is also good for us to feel pleasure[8].

I have also suggested that the most plausible version of the desire theory is a rationality-oriented (and awareness-oriented) Success Theory which contains certain "objectivist elements", viz. the following ones:

(iv) If a person has an intrinsic desire for a situation that is (in the prudential sense) "worth avoiding", then it is not good for him to have it fulfilled.

(vii) Desires for situations that are (in the prudential sense) worth desiring are more relevant than desires the objects of which are not (in this sense) worth desiring[9].

---

[7]But cf. note 6.

[8]But isn't this inconsistent with the relational idea that an experience is pleasant if and only if (and because) the subject has an intrinsic desire to have it when he has it? Doesn't this idea imply that all pleasant experiences are desire-fulfilments and that desire-fulfilment is (for this reason) the only type of thing that is good for us? Yes, it does, but only on the object interpretation of the desire theory; when the preference-hedonist attributes prudential value to a pleasant experience he does not attribute value to the circumstance that an intrinsic experiential desire is fulfilled, but to the object of the desire, viz. the experience itself. But since we have adopted the satisfaction interpretation of the desire theory, we must regard pleasure and desire-fulfilment as two different types of goods. That is, there is a sense in which it is correct to say that desire-fulfilment is all that matters, i.e. the following one: Every situation that has nonderivative value for a person is either a desire-fulfilment or the object of a desire.

[9]To get a better understanding of this type of combination between the desire theory and "the objective list theory", it is fruitful to assume the object interpretation of the desire theory, i.e. to regard the desire theory as a "subjective list theory", according to which a person's intrinsic preference ordering (subjective

That is,

P desires that X X obtains P believes that X X is on the positive "objective list"

is better for P than

P desires that X X obtains P believes that X X is neither on the positive nor on the negative "objective list"

which is (in turn) better for P than

P desires that X X obtains P believes that X X is on the negative "objective list"

(which does not have any positive prudential value at all).

In the chapters on the non-internalist pluralist theories, I claimed that there are no objective prudential values such that it is good for all of us to "possess" these things, and such that a person's well-being is a simple function of how much he possesses of them, regardless of whether we regard these "objective goods" favourably or unfavourably.

However, there are "objective prudential values" such that their presence make certain wholes more prudentially valuable than they would otherwise have been. Or more specifically, there are "objective prudential values" such that it is *ceteris paribus* nonderivatively better for a person to take pleasure in these things than to take pleasure in things which do not belong to this category, and such that it is *ceteris paribus*

---

list) coincides with his "value ordering". We can then see that there are two ways in which the existence of "objective prudential values" can affect this value ordering, namely (i) it can remove certain items from the positive subjective list, viz. the things which are worth avoiding, and (ii) it can improve the relative standing of certain items on the list, viz. it can (so to speak) move the things which are worth desiring up the list.

nonderivatively better for a person to have his desires for these things fulfilled than to have his desires for other things fulfilled (cf. pp 368-371 above)[10].

The theory itself

So, how should all these claims (and some other relevant claims) be combined (or "fused") into a coherent theory? Or more specifically, if we accept all these claims, how should our three central questions be answered? That is: **(I)** What kinds of situations are nonderivatively good for a person? **(II)** How do we determine just how (nonderivatively) valuable a certain (good) situation is for a certain person? For example, how do we compare different possible situations with respect to their nonderivative value for a certain person? And **(III)** How do we determine just how well off someone is at a certain time? For example, how do we compare different possible lives-at-certain-times with respect to their value for a certain person?

(I) There are two kinds of situations that are good for a person, viz. (a) to have certain kinds of pleasant experiences, and (b) to have his relevant intrinsic now-for-now desires fulfilled, but only on the assumption that he is aware of the objects of these desires.

That is, the answer to (I) is in part hedonistic and in part desire theoretical. "The objective list theory" enters the picture as follows: First, it is not good for a person to take pleasure in something that is on the negative objective list, and second, it is not good for a person to have a desire fulfilled if its object is on the same negative list.

(II) So, how do we determine just how good a certain (good) situation is for a certain person?

(a) In the case of valuable pleasures, the value that it has for a person to have such an experience is normally a function of how pleasant the experience is, e.g., it is better to have a more pleasant experience than to have a less pleasant experience. But sometimes, the prudential value of a pleasant experience is not just dependent on how pleasant the experience is, but also on other things, e.g., on whether it is based on

[10]I have not said very much about what these values are, however, i.e. the theory is rather incomplete at this point.

true or false beliefs. In particular, it is *ceteris paribus* better for a person to take pleasure in something that is on the positive objective list than to take pleasure in something that is not on this list (cf. pp 368-369 above).

(b) In the case of valuable desire-fulfilments, the value that it has for a person to have a relevant desire fulfilled is normally a function of how strong the desire is, e.g., it is better to have a stronger desire fulfilled than to have a weaker desire fulfilled. But sometimes, the value that it has for a person to have a relevant desire fulfilled does not just depend on how strong it is, but also on other things, viz. on whether or not the object of the desire is worth desiring (in the prudential sense). In particular, it is *ceteris paribus* better for a person to have an intrinsic desire fulfilled if its object is on the positive objective list than if it is not (cf. pp 370-371 above).

This part of the answer to (II) is in part hedonistic, in part desire theoretical, in part a combination between hedonism and "the objective list theory", and in part a combination between the desire theory and "the objective list theory". But how can (and should) the hedonistic theory and the desire theory be combined?

(c) This is one way in which we can compare the value of having a pleasant experience with the value of having a relevant desire fulfilled: The relational theory of pleasantness implies that the degree to which a pleasant experience is pleasant is a function of how strongly the experience is intrinsically desired by the experiencing subject. This suggests that we can (in principle) compare the value of a pleasant experience with the value of a desire-fulfilment, viz. by comparing the strength of the relevant desires, *but only if* the pleasure in question is a pleasure the value of which is dependent on its pleasantness alone and the desire-fulfilment in question is a desire-fulfilment the value of which is dependent solely on the strength of the desire. So, is this a good suggestion? Well, it is (in a way) hard to tell, since we do not (it seems) have any intuitions against which it can be tested. But on my view, the fact that the suggestion is very much in line with the desire theory gives it some support.

(d) There is yet another way in which the two theories can (and should) be combined, however: It goes without saying that it is (on this mixed theory) better to have a relevant desire fulfilled and to feel good about this than to have the desire fulfilled without feeling good about it. We can also assume that if a person is happy because he desires that

X and believes that X is the case, then it is better for him if X is actually the case, i.e. then it is better for him if the desire is actually fulfilled than if it is not. However, I also want to suggest that (1) the difference in prudential value between the whole

P takes pleasure in X
P desires that X
P believes that X[11]
X obtains

and the whole

P does not take pleasure in X
P desires that X
P believes that X
X does not obtain

is larger than (2) the difference in prudential value between

P takes pleasure in X
P desires that X
P believes that X
X does not obtain

and

P does not take pleasure in X
P desires that X
P believes that X
X does not obtain

(a difference which the hedonist regards as crucial) *plus* (3) the difference in prudential value between

P does not take pleasure in X
P desires that X
P believes that X
X obtains

and

[11]Here, and it what follows, I assume that P takes pleasure in X because he desires that X *and* believes that X. I also assume that the desire is relevant (e.g., that X is part of P's life).

P does not take pleasure in X
P desires that X
P believes that X
X does not obtain

(a difference which the desire theorist regards as crucial).

(e) The fact that it is (in all these cases) better for P if X is on the positive objective list than if it is not (cf. (II:a) and (II:b) above) suggests that all the three theories can (of course) be combined in a similar way. Or more specifically, (1) the difference in prudential value between

P takes pleasure in X
P desires that X
P believes that X
X obtains
X is on the positive "objective list"

and

P does not take pleasure in X
P desires that X
P believes that X
X does not obtain
X is on the positive "objective list"

is larger than (2) the "hedonistic" difference in prudential value between

P takes pleasure in X
P desires that X
P believes that X
X does not obtain
X is on the positive "objective list"

and

P does not take pleasure in X
P desires that X
P believes that X
X does not obtain
X is on the positive "objective list"

plus (3) the "desire theoretical" difference in prudential value between

P does not take pleasure in X
P desires that X
P believes that X
X obtains
X is on the positive "objective list"

and

P does not take pleasure in X
P desires that X
P believes that X
X does not obtain
X is on the positive "objective list"

In short, it is (roughly speaking) "very good" for a person P if complex facts of the following type obtain:

P takes pleasure in X
P desires that X
P believes that X
X obtains
X is on the positive "objective list"

(The view characterized by Parfit (cf. the quotation on p 367) is much stronger, though. On this view, it is *only* situations of this type which can have nonderivative value for a person).

(III) So, how does our mixed theory suggest that we determine how well off a certain person is (on the whole, and at a certain time)? Well, a person's well-being is (roughly) a function of how much valuable pleasure and how much valuable desire-fulfilment there is in his life. To capture the spirit of this idea, it is fruitful to formulate it (a) in terms of satisfaction, and (b) in terms of happiness.

(a) The term "satisfaction" often refers to the fulfilment of a desire, but it can also refer to the pleasure obtained from such fulfilment. If we let the term refer to both these things, i.e. to the pleasure obtained from the actual (and not just apparent) fulfilment of a desire, we can say that a person's level of well-being at a certain time is (roughly) a

function of how satisfied he is with his existence at that time.

(b) The term "happiness" often refers to a pleasant feeling, but the phrase "P is happy with X" can also mean that P takes a positive attitude toward X, that he endorses it, or regards it favourably. If we let the term refer to both these things, i.e. if we regard happiness as a mental state which consists of one affective and one attitudinal component (as a pro-attitude accompanied by pleasant emotions), we can say: A person's level of well-being is (roughly) a function of how happy he is with his existence, but only on the assumption that the affective component is based on true beliefs on what his existence is like.

Now, so far, this is very similar both to Sumner's (1996) subjectivist theory of welfare[12] and Nordenfelt's (1991) happiness theory of the quality of life[13]. My theory differ from their theories, however, viz. in that it recognizes certain "objective prudential values". So, how should the idea that a person's well-being is a function of how satisfied (or happy) he is with his existence be modified in order to account for this? Well, this is one way in which this can be done: The idea that there are certain "objective prudential values" suggests that there are a number of important dimensions (or domains) in a person's life, and that a list of the relevant values can be used to specify these domains (e.g., the social dimension). My objectivist idea can now be formulated as follows: A person's level of well-being is (roughly) a function of how satisfied he is with his existence, *but only as long as he takes all the relevant dimensions into account*. Or more specifically, the satisfaction which determines how well off a person is (on the whole) must (so to speak) "include" how satisfied he is in a number of "objectively pre-determined" areas, and a person's level of satisfaction in the relevant areas must (roughly speaking) be "in line with" the different objective values.

In short, to be well off is to be happy for the right reason.

---

[12]It is worth mentioning that there is more to this theory than what has been indicated this far. Sumner also requires that the endorsement of one's life (the attitudinal component of being satisfied with one's life) is authentic, or more specifically, that it (the endorsement) is both informed and autonomous.

[13]It can be doubted whether there is any genuine desire theoretical element in Nordenfelt's theory, however. On his view, there is no necessary connection between P's happiness and *the fact* that P desires are fulfilled; what matters for P's happiness is (rather) that P *believes* that his desires fulfilled.

Appendix A
How goodness-for differs from, and is related to, other kinds of value

The reason why I think it is important to contrast goodness-for (in particular, nonderivative goodness-for) against other kinds of values is mainly pedagogic. I know from experience that many philosophers (including moral philosophers) are not very familiar with the notion of goodness-for, and that it is (for this reason) not always easy to distinguish (nonderivative) goodness-for from a number of other "phenomena". The phenomena that have to be kept distinct from goodness-for (in general) are (a) "subjective value" and (b) agent-relative value, and the phenomenon that has to be kept distinct from final goodness-for is (c) final value-period.

There are at least two reasons for discussing how goodness-for is related to certain other kinds of values. First, it may contribute to a clearer understanding of what kind of goodness goodness-for (in particular, nonderivative goodness-for) really is, and second, it may throw some light on the issue of whether the notion of goodness-for is a normatively relevant notion, and if it is, in what way. The relations that are (on my view) most important to gain knowledge about are (d) the relation between goodness-for (in general) and "value-period" (in general), both agent-relative and agent-neutral value-period, and in particular, (e) the relation between final (e.g., nonderivative) goodness-for and final value-period.

So, let us now take a closer look at these contrasts and relations.

Goodness-for vs. subjective value

When we say that something is good for a certain person, or to a certain person, we sometimes mean that it appears to him as good, that it is good according to him, that he values it, or the like. Now, it is clear that something can be good for a person without appearing to him as good, and vice versa, and that it is (for this reason) important to distinguish between the two. In fact, *subjective value* (as e.g., Nagel (1986) and Hospers (1967) call it), or goodness-from-a-point-of-view (as Thomson (1992) calls it), is not really a kind of goodness at all. As Thomson points out, "a person's valuing something is an entirely subjec-

tive fact about that person and the thing he or she values, a subjective fact whose existence does not in itself imply that the thing valued is itself in *any* way good" (p 98).

Value-period

In order to be able to distinguish (nonderivative) value-for from things like intrinsic value, final value, and agent-relative value, we must first say something about *"value-period"* (or absolute goodness, or ethical goodness, or whatever one may like to call it). The reason why this "background work" has to be done is that the distinctions between intrinsic and extrinsic value, final and instrumental value, and agent-relative and agent-neutral value are (at least originally) all distinctions that are (so to speak) made within the realm of "value-period".

A second reason for paying some attention to value-period is this: Goodness-period is the kind of goodness that moral philosophers have paid most attention to, e.g., it is the kind of value that is considered most normatively relevant (it is, supposedly, the kind of goodness that utilitarians think we ought to maximize), and it also seems to be the kind of value that most of metaethics has been about (many metaethical theories of value, e.g., emotivism or internalism, are, on my view, most plausible if they are conceived of as theories about value-period)[1].

A third reason for paying attention to value-period is that the different theories of prudential value, e.g., hedonism or the desire-fulfilment theory, are normally conceived of as theories of "value-period" (or rather: as theories of intrinsic, or final, "value-period"). This makes it important to compare the notion of value-for with the notion of value-period, and to try to find out how the two types of values are connected to each other.

So, what kind of value is value-period? Or (I think) alternatively put: When the utilitarian claims that the right act is the act that maximizes the good, what kind of goodness does he have in mind? This is my answer: (1) To be good-period is to be good in the "predicative sense", (2) to be good-period is to be good "absolutely", or "to make the world a better place", (3) value-period is (supposedly) possessed by situations, and (4)

---

[1] This is not to say that all (or most) "truths" about value are really about value-period, however. Some truths, e.g., the idea that value is always a matter of degree, or the idea that values are supervenient, are supposedly truths about all kinds of values.

it cannot be ruled out that value-period is really a normative (or quasi-normative) notion, or that it must be understood in normative terms.

(1) To be "good period" is to be "just good", or (more precisely) good in "the predicative sense". So, what is it to be good in the predicative sense? Geach (1956) made the following distinction between predicative and attributive adjectives: "I shall say that in a phrase 'an A B' ('A' being an adjective and 'B' being a noun) 'A' is a (logically) predicative adjective if the predication 'is an A B' splits up logically into a pair of predications 'is a B' and 'is A'; otherwise I shall say that 'A' is a (logically) attributive adjective" (p 33). "Big" is offered as an example of an attributive adjective ("x is a big flea" does not split up into "x is a flea" and "x is big"), and "red" is an example of a predicative adjective (since "x is a red book" logically splits up into "x is a book" and "x is red").

He then went on to claim that "'good' and 'bad' are always attributive, not predicative, adjectives" (ibid., p 33). But if we look more carefully at what he says, we can see that he makes two distinct claims. The first claim is that all uses of the term "good" are what Beardsley (1981) calls adjunctive, i.e. that the term "good" is always (explicitly or implicitly) "adjoined, or affixed, to a noun or noun-phrase" (p 524). Or in Geach's own words: "Even when 'good' or 'bad' stands by itself as a predicate, and is thus grammatically predicative, some substantive has to be understood; there is no such thing as being just good or bad, there is only being a good or bad so-and-so" (p 34). The second claim is that "good" is always attributive, i.e. that statements of the form "X is a good K" never split up logically into "X is a K" and "X is good". Or in Thomson's (1994) terms, what Geach suggested was "that we should regard "good" /.../ as /.../ a predicate modifier: what is predicable of a thing is not (just) being good but rather being a good K, for some kind K", where the kind-term K is (so to speak) modified by the term "good" (p 8).

But there are many philosophers (like Ross and Beardsley) who thinks that Geach's claim is too strong. These philosophers claim that the term "good" can also appear by itself as a predicate, i.e. that statements of the form "X is good" are sometimes complete, and need not be spelled out as statements of the form "X is a good K". Or in Geach's own terms, "they allege that there is an essentially /.../ predicative /.../ use of the terms ['good' and 'bad', e.g.,] in such utterances as

'pleasure is good' and 'preferring inclination to duty is bad'" (p 35). He also adds that this (alleged) predicative use of "good" is often regarded as the philosophically important use of the term.

(2) Thomson (1994, 1996) also thinks that there is no such thing as being just good, or alternatively, that all statements of the form "X is good" are incomplete. But when she claims that there is no such thing as "pure, unadulterated goodness", she does not have Geach's view in mind. On her view, the adjective "good" is incomplete in a different sense than the one Geach had in mind, namely the following one: "[I]f Alfred says something of the form "X is good", then either he or the context must tell us *in what way or ways* he is saying the thing is good if we are to know what he is saying about it" (Thomson (1994), p 11, my italics). That is, when people say that something is good, "what they mean is always that the thing in question is *good in a way*, a way that the context of utterance, or the speaker, has to supply on pain of our simply not knowing what he or she does mean" (Thomson (1996), p 128). Alternatively put, "all goodness is goodness-in-a-way" (Thomson (1994), p 11), and "a thing's being good just *is* its being good in this or that way" (ibid., p 11).

So the question arises: Is goodness-period a way of being good? Well, not according to Thomson (1992) and (1996)[2], but she may be not be right about this. Consider the following three statements: "Ecological diversity is good", "it is better if 300 people die than if 700 people die", and "distributive equality is *ceteris paribus* better than distributive inequality". In my view, these are intelligible statements, and it is obvious that the kind of goodness (and betterness) that is being referred to in these statements is not goodness-for, but goodness-period. This suggests that *if* all goodness is goodness-in-a-way, and *if* the three statements are genuine evaluative statements, then we can conclude that goodness-period is goodness-in-a-way. (But as we will soon see (in (4)),

[2]Goodness-period is, for example, not among the ways of being good listed by Thomson (1992) and (1996). The ways of being good listed there are (1) the useful, or equipment goodness, or what von Wright (1963) calls instrumental goodness (to be good for use in doing something, or in achieving a certain purpose), (2) the skillful, or activity goodness, or what von Wright calls technical goodness (to be good at doing this or that), (3) the enjoyable, or aesthetic goodness, or what von Wright calls hedonic goodness (e.g., to be good to look at or listen to, to taste or smell good), (4) the beneficial, or goodness-for, and (5) the morally good (e.g., to be just, or generous, or tactful, or brave).

it may well be the case that goodness-period-claims are normative (or quasi-normative) claims rather than pure predications of goodness, and if this is so, we may have to reject the idea that being good-period is a way of being good)[3].

But suppose that goodness-period is goodness-in-a-way. We then have to ask: If something is good-period, in what way is it good? Well, the kind of goodness I have in mind here seems to be identical with what Korsgaard (1983) calls *"absolute goodness"*. That a thing is good absolutely means, in her terminology, that "here and now the world is a better place because of this thing" (p 169). In other words, to be good-period is (among other things) to make the world (as a whole) better[4]. But what kind of goodness (betterness) is it that is being attributed to the world here? Well, the answer is (of course) goodness-period, and this suggests that Korsgaard's idea is not very informative. And since it is also possible to attribute value-for to worlds, we can not define value-period as the kind of value that can be attributed to worlds (actual or possible).

(3) Many philosophers (including myself) hold the view that value-period is appropriately attributed to situations (or facts); in this respect, there is no difference between value-period and value-for. But there are exceptions to the rule. An example of this is Wetterström (1986), who claimed that the carriers of value-period are (ultimately) "concrete entities" (like things or parts of things, concrete events, or sequences and unities of such events) rather than situations. On this "entity-view", it makes perfect sense to conceive of things like pleasant experiences or knowledge as carriers of final value-period. The "situation-view" rejects this idea, and claims instead that value-period can (strictly speaking) only be attributed to things like the fact that someone has a pleasant

---

[3]It is worth mentioning that all this is based on the assumption that there is a certain kind of gap between the normative and the evaluative. For example, genuine predications of goodness (badness, or betterness) should, on this view, never be regarded as what Thomson (1996) calls *directives*, i.e. as claims "by the assertion of which we predicate of a person that he or she ought or should or must or is morally required, or is under a duty or obligation, to do or to refrain from doing, a thing" (p 131). There is a more detailed discussion of the normative/evaluative gap in appendix B.

[4]One could also say: That a good-period situation obtains is (in a manner of speaking) "good for the world", or "good from the point of view of the universe, or *sub specie aeternitatis*". This is somewhat misleading, however, since the world is (supposedly) not some kind of supersubject with its own perspective!

experience, or the fact that someone knows something. Notice that on my understanding of the situation-view[5], Wetterström's view is really the broader of the two views: Both views claim that concrete situations can be carriers of value-period, but Wetterström also claimed that there are other kinds of concrete entities that can possess value-period[6].

In the following, it will be assumed that the carriers of value-period are situations. But what is it for a situation to be good-period? And what is it for one situation to be better-period than another one? All we know at this point is that a situation is good-period if it makes the world better-period than it would otherwise be, and that one situation is better-period than another situation if it makes a larger contribution to the "goodness-period of the world". But what else is involved in the notion of value-period?

(4) The only plausible answer to this question is (I think) the idea that value-period must somehow be understood "in terms of" what we *have a reason to* do (try to achieve) or want or like, or what it is *rational* to do or want, or what we *should* do or want, or what we *ought to* do, or the like[7]. We should also add that the reasons (etc.) in question are conceived of as *objective*, e.g., to say that something is good-period is to say that we have a reason to promote it (that it is rational for us to promote it, that we ought to promote it, etc.), *regardless of what we actually want and believe* (cf. Nagel (1986)). To be a little more precise (but not sufficiently so):

---

[5]Where particular facts are regarded as concrete entities, rather than as (true) propositions, or the like (cf. chapter 1, note 4). Cf. also note 6 below.

[6]It can also be argued that value-period can also be carried by *types of situations* (or "propositions"). This does not seem like a plausible view, however, but this does not mean that the idea that the carriers of value-period are concrete situations (rather than types of situations) is unproblematic. The problem is that the view is based on the conception of value as property, but what happens if we drop this view? If the goodness-period is, instead, understood in terms of reasons to promote (cf. (4) below), it would seem that all "strict talk" about carriers of value is nonsensical. We could even say (in a more figurative sense) that it is types of situations that are the prime carriers of value, since what should primarily be promoted is that some *type* of situation is realized.

[7]Where "we" does not necessarily mean "everyone": it can also refer to me, or you, or some of us. That is, I do not accept the idea that "[a] state of affairs is good absolutely if and only if for all A, other things being equal, A ought to aim at it" (cf. Thomson (1992), p 113). Or alternatively put, I want to allow for the possibility that there is such a thing as *agent-relative* value-period (cf. below). (Whether there is such a thing as *intrinsic* agent-relative value-period is (of course) an entirely different matter).

That a certain situation is *good-period* either "means" (in a rather loose sense of the term) (i) that we have a reason to want it to happen, that it is desirable, that it is worth desiring (liking), or the like, or (ii) that we have a reason to pursue it, promote it, aim at it, try to achieve it, produce it, or preserve it, or that it is worth promoting, pursuing, or aiming at.

That a certain situation is *bad-period* either "means" (i) that we have a reason to want it to stop, or that it is rational to want to avoid it, or (ii) that we have a reason to try to get rid of it, that it should be got rid of or avoided.

That a certain situation X is *better-period* than another situation Y either "means" (i) that X is to be preferred to Y, or that it is more desirable than Y, (ii) that if there is an exclusive choice between X and Y, then we should choose X.

Now, it is tempting to understand the reasons to want and like (the (i)-clauses) in terms of reasons to do (the (ii)-clauses). The only reason why certain things are worth wanting (it might be argued) is that there is a conceptual link between desire and action, i.e. if we would never act on our desires, we would not have a reason to desire anything. This suggests that there can not be a reason to want X unless there is a reason to promote X (in action). It would, for example, be implausible to claim that X is worth wanting but not worth pursuing[8].

But how should a (vague) expression like "A has a reason to promote X" be understood? As I see it, it can either be understood as (A) "it is prima facie rational for A to promote X" (i.e. in terms of rationality), or as (B) "A has a prima facie (moral) duty to promote X", or "to promote X is the morally right thing for A to do" (i.e. in terms of morality)[9]. In this context, there is no need to choose one of the options, however. Suffice it to say that to choose (A), as Kagan (1989) seems to do, is to analyse the good in terms of the right. Such an analysis makes phrases

---

[8]It is worth noticing that this line of argument treats the cases where the very wanting of a certain thing has beneficial effects as negligible.

[9]This is why the phrase "prima facie" has to be included: What we have most reason to do (e.g., promote) in a particular case ("all things considered") depends on what the alternatives are, i.e. we sometimes have a reason to promote the lesser of two evils, and we do not always have a reason to promote what is good (viz. if there is a possible outcome that is better). That is, we do not always have a reason (*in this sense*) to promote the good, or to try to get rid of the bad; therefore "prima facie". There is another way to deal with this problem, however, viz. to understand both "good" and "bad" in terms of "better" (cf. appendix B).

like "we ought to promote the good" become rather empty, or as Kagan himself writes: "to say that there is a pro tanto reason to promote the good is actually to make a trivial claim" (p 61)[10].

Another source of unclarity is this: How exactly should the idea that goodness-period should be "understood" in terms of reasons to promote (etc.) be understood? For example, should it be understood as (a) "The claim that X is good-period has the same meaning as the claim that we have a prima facie reason to aim at X", or as (b) "X is good-period if and only if we have a prima facie reason to aim at it", or as (c) "If X is good-period, then this is (by definition) something we have a prima facie reason to aim at, but not vice versa"? Well, on my view, we should choose (b) (which is, I think, more intuitively plausible than both (a) and (c)).

Now, *if* this analysis is correct, then the notion of value-period is not merely "extremely normatively relevant"; there is also a good reason to regard it as a *normative notion*, rather than as a genuinely evaluative notion (whatever *that* is). This would not just "explain" why the claim that there is a reason to promote the good is a trivial claim; it would also imply that we can not really say that we should aim at X *because* it is good-period.

This point is intimately connected to (or even identical with) another point, namely: If the analysis above is correct, there is reason to believe that goodness-period can (or should) not (unlike the other ways of being good) be regarded as a property[11]. Suppose that goodness is

---

[10]It is, however, important to point out that this analysis does not make utilitarianism empty. The claim that we always ought to maximize the good is much stronger than the claim that we have a prima facie duty to promote the good; the utilitarian claim also involves the idea that there are no other duties which may override the duty to promote the good.

[11]Or more specifically, that *intrinsic* goodness-period can not be regarded as a "property" in the metaphysical sense, i.e. as an "intrinsic property", or as "something "in" the object", "possessed by the object", or "carried by the object". The term "property" can also be used in a "semantic sense", however, viz. as synonymous with "predicate". The semantic term "property" is intimately connected to the notion of truth: Here, the statement "X has the property P" is regarded as equivalent with the statement "It is true that X is P". That is, to reject the idea that goodness is a property in this sense is to reject the idea that value-statements can be true. But if the idea that goodness is not a property is (instead) regarded as a metaphysical claim, then it is compatible with the idea that value-statements can be true.

It is also worth noting that the idea that intrinsic goodness-period is not a property does not really constitute any threat to the important idea that value-

conceived of as to-be-pursuedness, badness as to-be-avoidedness, and betterness as to-be-preferredness. Now, these are indeed "metaphysically mysterious", or even "queer" properties (cf. Thomson (1996), pp 127-128). So why regard them as properties at all? I think this makes it easy to understand why so many philosophers have been attracted to metaethical theories like emotivism, prescriptivism, nihilism, and internalism. That is, if goodness is regarded as to-be-pursuedness, it is easy to understand how one can come to believe that "good" is a commending word, that sentences of the form "X is good" lack truth-value, and so on[12].

The idea that goodness-period should be understood in terms of reasons to promote is rather problematical, however. For example, if "ought" (e.g., "having a prima facie reason to promote") implies "can", then this analysis seems to imply that unrealizable situations cannot be good-period, but this is surely counter-intuitive. Is this an insurmountable problem? Well, not necessarily. First, we need not regard the idea that "ought" implies "can" as a *conceptual* truth, and second, we can always incorporate "can" in our analysis: for example, we can always claim that X is better-period than Y if and only if it is the case that if some agent could realize X instead of Y, he should do so (this is Tännsjö's proposal)[13]. Another problem with the suggested analysis is this: As we all know, it is possible to adopt a normative theory that is purely deontological, i.e. that does not contain any axiology at all. Now,

---

period is supervenient (cf. appendix B).

[12]To drop the conception of goodness as property has an interesting (and desirable) consequence: It becomes much easier to understand why we can attribute value to hypothetical situations, e.g., to possible lives, and it becomes possible to avoid a number of intricate questions about value and time, e.g., questions of the form "when does this-or-that have value?", or questions like "can a situation have value before it obtains, or after it stopped obtaining?".

[13]But if this manoeuvre is blocked, what should we do? Should we (instead) try to analyse value-period in terms of what we have reason to want or like (cf. the (i)-clauses on p 385 above)? Well, it seems that such an analysis would not be open to the objection above; it is, after all, not very plausible to assume that the (i)-clauses can be understood in terms of the (ii)-clauses; in particular, it is implausible to assume that reasons to like can be "reduced to" reasons to do. However, as it stands, this analysis fails to distinguish value-period from other kinds of values, e.g., aesthetic value, and it must (therefore) be supplemented by some idea of what kind of wanting or liking we have in mind when we say that something is good-period if and only if we have a reason to want or like it. This is no easy matter, however. Perhaps we have to resort to the "Moore-like" view that value-period is *sui generis* and unanalyzable.

as I see it, it is quite possible that such a pure deontologist believes that certain value-period-statements are true. And if this is the case, doesn't this imply that value-period-statements can not be regarded as directives after all? Or should we (instead) claim that our deontologist does not *really* understand what it means that something has value-period, or that he doesn't *really* believe that there are true value-period-statements?

This ends the general characterization of value-period. We are now in a much better position to grasp the three distinctions that were mentioned above. The first of these distinctions was the one between intrinsic and extrinsic value-period.

Intrinsic vs. extrinsic value-period

The intrinsic goodness of a thing is (roughly) the value it has "in itself", and the extrinsic goodness of a thing is the value it gets (or "derives") "from some other source" (cf. e.g., Korsgaard (1983), p 170). On this view, the difference between intrinsic and extrinsic value is conceived of as a difference in "location" of value, or "source" of value: and to say of something that it is intrinsically or extrinsically good is to make a claim about the "location" or "source" of its goodness.

This can be interpreted in at least two different ways. On the first interpretation (suggested by Thomson (1992)), something is intrinsically good just in case it does not derive its value from somewhere else[14], i.e. intrinsic goodness is conceived of as nonderivative goodness[15]. On the second interpretation, something is intrinsically good if and only if it is good independently of all other "objects", or alternatively: if it is good in all possible worlds, or if it is good in virtue of its intrinsic nature, or if its goodness supervenes on its intrinsic features, or the like[16].

---

[14]From the value of something that is distinct from it, that is. Wholes can have intrinsic value (cf. the discussion on p 7).

[15]This is probably what Wetterström (1986) had in mind when he said that an evaluation is intrinsic in an ethical system only if it is independent of other evaluations in this system.

[16]This is the notion of intrinsic value that Kagan (1992) adopts. That an object has intrinsic value means, on his view, that it has value "independently of all other objects", and the intrinsic value of a thing is conceived of as "the value it would have even if it were the only thing existing in the universe. If anything does indeed have value in this sense, then it seems clear that such intrinsic value must depend solely upon the intrinsic properties of the object" (p 183).

It is worth adding that there is a certain "tension" in all this. The idea that the

It is not hard to see that these are two different interpretations: To say that the goodness of a thing is derivative is (roughly speaking) to say that it inherits its goodness from some other good thing, and if we keep this in mind, it is pretty obvious that the question of whether a thing is good in virtue of its relational properties or not has little or nothing to do with the question of whether its goodness is derivative or not. In particular, a thing's being good in virtue of its relational properties does not imply that its goodness is derivative. So, having pointed out the difference between these two senses of "intrinsic value", let us reserve the term "nonderivative value" for the first one and the term "intrinsic value" for the second one.

This notion of intrinsic value is based on the idea that the goodness of a thing is something that supervenes on so-called natural features (intrinsic or relational) of this thing, or alternatively put, it is (as I see it) based on a conception of value as a supervenient *property*. But as we have seen, the conception of goodness as a property is not easy to harmonize with the conception of goodness as to-be-pursuedness. This does not mean that the notion of intrinsic to-be-pursuedness is a nonsensical notion, however (or that it is based on a "category-mistake"). To say that a certain situation has "intrinsic to-be-pursuedness" is simply to say (i) that we have a prima facie reason to promote it, and (ii) the reason why we have a reason to promote it is that it has certain intrinsic features, and this makes perfectly good sense.

But the fact that the notion of "intrinsicality" makes sense (even when value is understood as to-be-pursuedness) does not imply that it is a normatively relevant notion. In fact, I would like to suggest that this notion lacks all normative relevance. First, it seems plausible to assume that the notion of value-period is normatively relevant only if value is understood as to-be-pursuedness. But the fact that a certain situation is *intrinsically* good in this sense is (I think) not more normatively relevant than the mere fact that it is good. Or alternatively put, it is of no nor-

---

intrinsic value of X is the value it has in virtue of its intrinsic properties, on the one hand, makes most sense if X is a thing rather than a situation. (On my view, it makes much more sense to talk about (correct) descriptions of situations than to talk about properties of a situation. But is a situation really distinct from its descriptions? If it isn't, it seems rather superfluous to say about a situation that it is good in virtue of its intrinsic properties). The idea that X is intrinsically good if and only if it is good in all possible worlds, on the other hand, makes more sense if X is a situation (or fact).

mative interest at all to find out whether a situation that is valuable in this sense is also intrinsically valuable in this sense. What a teleological moral philosopher needs to know is what kinds of situations we should promote, and why, and the idea of intrinsicality does not really throw any light at all on these questions[17].

Final vs. instrumental value

Now, this seems to contradict a very common view in moral philosophy, namely the view that intrinsic value is the most fundamental kind of value, and that it is therefore of utmost importance to find out what kinds of situations that have intrinsic value. Frankena (1973) writes:

> /.../ [I]n order to come to a judgement about whether something is good on the whole or good in any of the other senses, we must first determine what its intrinsic value is, what the intrinsic values of its consequences or of the experiences of contemplating it is, or how much it contributes to the intrinsically good life. Our task, therefore, is to determine the criteria or standards of intrinsic goodness and badness (p 83).

Now, what Frankena has in mind is obviously not intrinsic value in the sense it was given above, but rather what Korsgaard (1983) calls "final value", or "value as an end", a kind of value that is to be contrasted against "instrumental value" or "value as a means".

On Korsgaard's view, the difference between final and instrumental value is a difference between what can be called "ways of actually valuing a thing". To be good as an end is to be "valued for its own sake", and to be good as a means is to be "valued for the sake of something else" (p 170). But Korsgaard is mistaken in believing that this is a difference in value, her distinction is (rather) a distinction between two different ways in which something can be desired (or valued). To get a distinction between kinds of (objective) value, we need to replace "valued" with "valuable", "worth valuing", or the like. If we do this, we get: That something has final value (is valuable as an end) means that it is worth valuing (or desiring) for its own sake, or that it deserves to be valued (desired) for its own sake (cf. Thomson (1992), p

[17]For this reason, I will not bother to investigate how the notion of nonderivative goodness-for is related to the notion of intrinsic value.

108, and Kagan (1992), pp 183-184)[18].

This suggests that the notion of final value is based on a conception of goodness as desirability, rather than (as in the case of intrinsic value) on a conception of goodness as property[19]. (This is one reason why the common distinction between intrinsic and instrumental value is, as Korsgaard expresses it (on p 170), "misleading, a false contrast"[20]). This follows from the idea that the phrase "final value" has to be understood in terms of "worth desiring for its own sake", which is (in turn) "derived" (or "drawn") from the notion of "desiring for its own sake". (This would explain why Korsgaard made the mistake she did).

The fact that the notion of final value is (in this way) based on the conception of goodness as desirability means that it makes perfectly good sense to talk about final value-period. It also explains why the notion of final value is (to the extent that goodness is understood as to-be-pursuedness) of the highest normative relevance. If we assume that the question of final value-period is really a question about what kinds of situations that should be pursued as ends, then we can see why it is such a fundamental question in normative ethics[21]. It should also be

---

[18]Anders Tolland has pointed out (in conversation) that the phrase "X is worth desiring for its own sake" is ambiguous, viz. that it can also be understood in an instrumental sense, i.e. as "If you desire X for its own sake, this will have good consequences for you". Needless to say, this is *not* the sense in which the phrase should be understood.

[19]The notion of final value is not just in harmony with the conception of goodness as desirability, however, but also with the conception of goodness as to-be-pursuedness. (I assume that desirability and to-be-pursuedness are two different things; cf. above). All this is very much in line with the fact that it makes more sense to attribute final value to situations than to things.

[20]The main reason why it may be misleading to make this distinction is (of course) that the terms are "drawn" from two different distinctions, viz. the final/instrumental and intrinsic/extrinsic distinctions. It is important to keep these distinctions separate, however, and the main reason for this is that they may not be "extensionally equivalent" with each other. In particular, separating the two distinctions makes us see that something may be both extrinsically valuable and valuable as an end (cf. Korsgaard (1983), p 172). Or as Kagan (1992) puts it, there is, "on the face of it", "no reason to assume that what has /.../ value as an end /.../ must have this value solely by virtue of its intrinsic properties", and being valuable as an end does not entail being intrinsically valuable (p 184). One might even "consistently hold that absolutely nothing has intrinsic value /.../, while still insisting that many things have" final value (p 184).

[21]It is (of course) final to-be-pursuedness that most utilitarians thinks we ought to maximize. And if we instead believe (more modestly) that we have a "mere" prima facie duty to promote certain types of situations, then we must also, in order to find out what these situations are, know what is worth promoting as an

pointed out that it is very likely that all situations that are nonderivatively good-period are also finally good-period (but not vice versa, since it is possible that the final value of a whole is derived from the final values of its parts; cf. the discussion on p 7 above)[22]. And since attributions of nonderivative value are (by definition) "basic" in ethics, we can also say that evaluations of the form "situations of type X have final value-period" (where X is not a whole) are (in a sense) fundamental in an ethical theory.

Agent-neutral and agent-relative value

Above, we have looked at two of the different possible senses of the phrase "X is good for P", namely "That X holds is in P's interest" and "X is valued by P". There is one more interpretation to consider, however, namely "X has agent-relative value in relation to agent P".

Like the distinction between final and instrumental value, the distinction between agent-neutral and agent-relative value only makes sense if we conceive of value as to-be-pursuedness or desirability. So, let us assume that "X has positive value-period" "means" that one has a reason to want X to happen, or that one has a reason to promote X, where these reasons are objective, i.e. independent of one's actual desires and beliefs. Let us now assume that someone says about something that it has (objective) value in this sense. It is then legitimate to ask *who* it is that has an objective reason to promote it. Is it anyone, or is it just some particular person? It is in relation to this question that Nagel's (1986) distinction between agent-neutral (impersonal) and agent-relative (personal) objective value should be understood. To say that something has *agent-neutral value* is (roughly) to say that anyone has a prima facie reason to promote it (if it is positive) or try to stop it (if it is negative), and to say that something has *agent-relative value* is (roughly) to say that only some people have a reason to promote it, or more precisely: A situation has *agent-relative value in relation to agent P* if and only if P has a prima facie objective reason to promote it (if it is positive) or try to stop

---

end. The question of the finally good (where goodness is understood as to-be-pursuedness) is the perhaps most important evaluative question in moral philosophy.

[22]We could also add that all instrumentally good things are derivatively good, but not vice versa (to be good as a means is only one of several ways in which something can be derivatively good).

it (if it is negative)[23].

It seems that many of the things that have agent-relative value also have agent-neutral value. As an example, consider me being in pain. I surely have a prima facie objective reason to avoid my own pain: the fact that I am in pain has negative agent-relative value in relation to me. But supposedly, anyone has a reason to want any pain to stop, including mine: the fact that I am in pain is also bad in the agent-neutral sense. That is, that it has negative agent-neutral value to be in pain means that we do not just have a reason to avoid our own pain; we also have a reason to relieve the pain of others (at least when we are in a position to do so).

This gives rise to the following question: Are there objective values that are "merely agent-relative", i.e. which do not correspond to any agent-neutral values (reasons)? Nagel (1986) thinks so, and I think he is right. The central idea here is that there may well be interests and desires which do not generate impersonal values, but which give rise to personal values. To borrow an example from Nagel (p 167), suppose that someone has a strong desire to climb the top of Kilimanjaro. In this case, his reaching the top may well have relative value for him: he may have a good (objective) reason to try to get to the top. But other people may have very little reason, if any, to care whether he climbs the mountain or not. In this case, the fact that he reaches the top has relative value for him, but there is no corresponding neutral value. And if this example is not convincing enough, consider the following one: Suppose that someone has a desire that everyone is converted to Christianity. It is not implausible to assume that this (altruistic) desire gives him an objective reason to try to convert people to Christianity, i.e. that it generates a relative value for him. But this relative value hardly corresponds to any neutral value. In short, we do not have any good reason to assume that all agent-relative values correspond to neutral values.

We are now familiar with the relevant distinctions, and it is time to start looking at the rest of the contrasts and relations that were listed in the very beginning of this appendix, viz. (b) the contrast between goodness-for (in general) and agent-relative value, (c) the contrast between final goodness-for and final value-period, (d) the relation

---

[23]At this point, it might be fruitful to reread note 9 on p 385.

between goodness-for (in general) and agent-relative value[24], and (e) the relation between final goodness-for and final (agent-neutral) value-period. I will first look at the contrast and relation between goodness-for (in general) and agent-relative value, and then move on to the more central issue: How does final value-for differ from final value-period, and how are the two related to each other?

Value-for and agent-relative value

So, how is the notion of value-for related to the notion of agent-relative value? In particular, does the fact that certain situation has agent-relative value for a certain person imply that it is in his interest that this situation obtains, and vice versa?

Let us first point out that something may well have agent-relative value in relation to a certain person without having value-for-him. That everyone is converted to Christianity has relative value for the person in the example above, no matter whether this is in his interest or not. Or in more general terms, everyone has desires, projects, commitments, personal ties, and the like, that give him a reason to act in the pursuit of ends that are his own, regardless of whether this is in his interest or not.

So, does the fact that something is in a person's interest imply that it has agent-relative value for him? Or more specifically, does the fact that a situation X is better for me than another situation Y give me a prima facie reason to promote X? I think it does. If something is better for me, then I have a reason to choose it when there is a choice (and I have a reason to choose *because* it is better for me). On this view, everyone has a relative reason to promote his or her own welfare, but notice that this reason may often be overridden by other kinds of considerations, e.g., the fact that we also have a neutral reason (or prima facie duty) to promote the welfare of all sentient beings, or a relative reason to treat other people in a decent way, and so on. (The self-interest theory (or ethical egoism) only gives us a small part of the truth).

To sum up, the notion of agent-relative value and the notion of value-for are two different notions. However, value-for-claims can constitute reasons for claims about agent-relative value, but not vice versa.

---

[24]The only interesting "aspect" of the relation between value-for and agent-neutral value is the relation between *final* value-for and *final* agent-neutral value. However, this relation will be discussed in (e).

Final value-for and final value-period

Let us first note that the theories of the good life that will be discussed in this book (e.g., hedonism or the desire theory) are often conceived of as theories of final value-period[25], rather than as theories of prudential value. To illustrate the difference between the two, let us take hedonism as an example. Viewed as a conception of prudential goodness, the hedonistic theory claims (roughly) that the only thing that is ultimately in our interest is to have pleasant experiences. Viewed as a conception of final value-period (to-be-pursuedness), on the other hand, hedonism claims that the only thing we (ultimately) have a reason to promote is (roughly) that people's lives (or the lives of sentient beings) are as pleasant as possible. These are two very different claims, but they are intimately connected to each other. But how?

Or more precisely, how are (valid) statements of the form "It has final (e.g., nonderivative) value for a person P that X obtains" related to (valid) statements of the form "It has final (e.g., nonderivative) value-period that X obtains"? For example, are all valid statements of the latter type (e.g., "it has final value-period that P feels pleasure") "entailed by" valid statements of the former type (e.g., "it has final value for P that he feels pleasure")?

The first question is this: If it has final value for people that situations of a certain type obtain, does this imply that ("other things being equal") it has also final value-period that situations of this type obtain? Or are there situations that have final value for people without having final value-period?

In answer to this question, most philosophers would probably claim that if all situations of a certain type of have final value-for-people, then they also have final value-period. Or alternatively put, final value-period is (at least in part) a function of final value-for. On this view, there is no situation-type X such that (i) it is in a person's ultimate interest that situations of type X obtain, and (ii) it does not have final value-period that situations of type X obtain.

Moreover, the fact that a certain situation is good for someone normally constitutes a *reason* for regarding it as good-period, but not vice

---

[25]Where value-period is probably conceived of as to-be-pursuedness (this is what I will assume in the following), but where it may also be conceived of as desirability.

versa. For example, if we accept the plausible premise that human welfare has final value-period, it can plausibly be argued that the *reason* why it is finally good that people have pleasant experiences is that it is nonderivatively good for people to have pleasant experiences. In this way, hedonism *qua* conception of prudential goodness can constitute a reason for hedonism *qua* conception of final value-period[26].

But is it really certain that there is no situation-type such that situations of this type (i) have final value-for-people, but (ii) lack final value-period? As far as I can see, it is possible that there exist such situation-types, and what I have in mind here is desire-fulfilment. Suppose that it has final value for a person that his desires are fulfilled. Does this really "imply" that there is an "ultimate" reason to aim at the fulfilment of his desires? If we take the distinction between agent-relative and agent-neutral value-period into account, we can see that the following view is far from absurd: "The fact that it has final value for a person that his desires are fulfilled gives *him* a reason to try to fulfil his desires, but it does not give *the rest of us* any reason to try to fulfil *his* desires. We have already seen that there are agent-relative values that do not correspond to any agent-neutral values, so why not add that there are *final* relative values that do not correspond to any *final* neutral values, viz. certain kinds of desire-fulfilment? In short, it is not irrational to accept preferentialism *qua* conception of prudential goodness and (at the same time) reject preferentialism *qua* conception of final (agent-neutral) value-period altogether".

On this view, there is a situation-type that (i) has final value for people, but (ii') lacks final *agent-neutral* value-period, but there is no situation-type that (i) has final value for people, and (ii'') lacks final *agent-relative* value-period. This means that the (plausible) idea that human welfare should always be promoted is only partly rejected[27].

The second question is this: Is all final value-period value for someone, or are there types of situations that have final value-period without being (finally) valuable for anyone? Or alternatively put, is

---

[26]However, hedonism *qua* conception of final value-period does not *follow* from hedonism *qua* conception of prudential goodness unless what Rabinowicz calls axiological individualism is true, i.e. unless *all* final value is value for someone (cf. Rabinowicz and Österberg (1996), p 11). Cf. also below.

[27]More precisely, the idea is that preferentialism is plausible *qua* axiological component of egoism, but that it is implausible *qua* axiological component of utilitarianism.

"axiological individualism" true or false? There are two different ways in which this question can be interpreted: (a) Does every situation that has final value-period also have final value for someone (anyone?), e.g., for some existing person? (b) Is the final value-period of a certain situation a *function of* the final values-for which it "contains"? (As we will soon see, it is (b) that is the central question).

The answer to (a) is obviously "no". A situation may well have positive value-period without being in anyone's interest: That a certain situation has final value-period (or that a certain situation is finally better-period than another) does not imply that there is any particular (existing) person for whom it is good (or better). To see this, consider the idea that it is better that 100 babies are born than that 50 babies are born, or that it is better if 300 people die than if 700 people die. Here, it is quite clear that there need not exist any person (or sentient being) P such that "better" can be replaced by "better-for-P".

The central question is rather (b): Is the value-period of a situation a function of the values-for which it "contains"? Or alternatively: Are all final values-period "personal values", or are there also "impersonal final values" (e.g., distributive equality or ecological diversity)?

The axiological individualist claims that all final values are personal values, or more precisely: The value-period of a *particular* situation is a function of the values-for that it "contains", and this suggests that a (particular, actual) situation cannot have final value-period unless there is someone for whom it has final value. Notice that this view is compatible with the idea that the more people there exist, the better-period. The reason for this is that once a person has begun to exist, it is (in the vast majority of cases) good for him to continue existing.

However, axiological individualism rules out the possibility that there are truly impersonal values, and this means that if one such impersonal value can be found, then this view must be rejected. And if axiological individualism is false, then it might well be plausible to accept hedonism (etc.) *qua* conception of prudential goodness and reject hedonism (etc.) *qua* conception of final value-period.

So are there any truly impersonal final values, e.g., does ecological diversity, or distributive equality, have final value-period? There will be no discussion of this question in this book. My primary concern is value-for rather than value-period, and there will (for this reason) be no "direct" discussion of what one, *qua* utilitarianist, should try to maxi-

mize. It is true that an answer to this question (i.e. the question of value-period) would probably have to *include* a conception of prudential value, but such a conception could never constitute a complete answer to the question of final value-period (not even on the assumption that axiological individualism is true; we would also have to know how the number of welfare subjects which are included in a situation (or "distribution") affects the value-period of this distribution)[28].

[28]Notice that even if the question of final value-period and the question of final value-for would have (roughly) the same answer, this would not imply that the notion of final value-period and the notion of final value-for are identical. Again, the fact that situations of a certain type have final value for people can "explain" why it has final value-period that situations of this type occur, but not vice versa. Or in other words, the *reason* why it has final value-period that situations of a certain type obtain is that it has final value-for-people that situations of this type obtain.

Appendix B
Some further characterization of value-for

For those who want to gain a fuller understanding of the notion of value-for, it is worth looking at three more features that it has (or seems to have), viz. the following ones: (i) Even though the notion of value-for is a normatively relevant notion, and even though a conception of value-for is action-guiding; value-for is not a normative notion, and a conception of value-for is not directly or actually action-guiding; (ii) betterness-for is primary to both goodness-for and badness-for, i.e. goodness-for is fundamentally comparative; and (iii) value-for is (like all value) supervenient, but the way in which it is supervenient is different from the way (or ways) in which other kinds of values are supervenient.

Normative relevance and "action-guidance"

In this section, I will make two major claims, one claim about the "normative status" of value-for, and another claim about to what extent, and in what way, conceptions of value-for can be regarded as action-guiding.

The first claim is that even though the notion of value-for is a normatively relevant notion, it is not a normative (or quasi-normative) notion. Or alternatively, value-for-statements are "evaluatives" rather than "directives", and *qua* evaluatives, they do not in themselves tell us what we have reason to promote: to do this, they must be combined with some directive or other. In short, the claim is that there is a gap between the evaluative and the normative, and that value-for belongs on the evaluative side.

The second claim is that even though conceptions of prudential value are (in a sense) "essentially action-guiding", this does not mean that they are (in any sense) directly action-guiding, or that they are even "actually action-guiding". On the contrary, such conceptions are (rather) "indirectly" and "potentially" action-guiding. However, it is important to notice that this claim does not invalidate the point that the main reason why the question of prudential value is important is that

conceptions of prudential value are (essentially) action-guiding[1].

Let us now take a closer look on these two claims.

The normative relevance of value-for

The first claim can be divided into two parts; (A) one negative claim, viz. the idea that the notion of value-for is not a normative (or quasi-normative) notion, and (B) one positive claim, viz. the idea that the notion of value-for is a normatively relevant notion. Let us first look at the negative claim.

(A) As we have already seen (in appendix A), it is problematical but not absurd to regard the notion of value-period as a "quasi-normative" notion. The idea that claims of the form "X is good-period" are normative claims (or what Thomson (1996) calls "directives") rather than evaluative claims (or what Thomson calls "evaluatives") may not be valid, but it is at least worth taking seriously.

In Thomson's (1996) terminology, evaluatives are "sentences by the assertion of which we predicate goodness in a way, or badness in a way, or betterness in a way[2]". Directives, on the other hand, are "sentences by the assertion of which we predicate of a person that he or she ought or should or must or is morally required, or is under a duty or obligation, to do or to refrain from doing, a thing" (pp 130-131)[3]. Now, this seems to mean that *if* things like "promoting" and "trying to bring about" are included in the category "doings", then it is

---

[1]These are two different claims, but they are (I think) not always clearly distinguished from each other. But do we really need to distinguish between the claims? Yes, we do, and this is why: The two claims are extensionally equivalent only if the following two statements are true: (i) Norms (or directives) are directly action-guiding, i.e. "norm-internalism" is true, and (ii) the only way in which a conception of prudential value can be action-guiding is by being normatively relevant, or alternatively, nothing but norms is directly action-guiding. But (ii) is obviously false (a conception of value-for can also guide our choices in "non-normative" ways, e.g., by "hooking into" our actual desires), and (i) may well be false too (it is far from obvious that norm-internalism is true), and it is (therefore) important to keep the two claims above separate.

[2]Or, I would like to add, neutral value in a way!

[3]Personally, I think the distinction between the evaluative and the normative is best regarded as a distinction between different kinds of *claims* (or judgements, or the like) rather than between different kinds of *sentences*. After all, it seems that one and the same sentence (e.g., "this is good" or "this is right") can be used both normatively (i.e. to make a normative claim) and evaluatively (i.e. to make an evaluative claim).

at least not obvious that value-period-claims should not be regarded as directives. Value-for-claims, on the other hand, are "pure" evaluatives, i.e. there is no way in which they can be regarded as normative claims[4].

The idea that value-for-claims are not directives (which is but a special case of the more general idea that there is a conceptual gap between the evaluative and the normative[5]) can be illustrated as follows: The (evaluative) claim that something is good for a certain person does not imply that anyone (not even the person himself) has a reason to promote it, e.g., the claim that it is good for John to have pleasant experiences does not imply that John himself or anyone else should try produce pleasant experiences in John[6]. However, the fact that something is in John's interest surely gives John an objective reason to promote it (the reason why he has a reason to promote it is that it is good for him, he has a reason to promote it *because* it is good for him), and sometimes it gives other people a reason to promote it too. But the reason why we tend to accept this view is that we tend to regard directives like "everyone has a reason to promote his own welfare" or "if the costs are not too high, everyone has a strong (moral) reason to promote other people's welfare" as valid[7]. As far as goodness-for is concerned, there is a gap between "good" and "ought", and this explains why it is not empty (or trivial) to say that we ought to promote people's welfare.

That is, to say that a certain action would benefit a certain person, and harm no one else, but that one should refrain from performing it, this is not to make an inconsistent claim. Or to stretch the point even further, one may even consistently claim that a certain situation ought to be pursued just *because* it is bad for a certain person. But it is impor-

---

[4]As I see it, this is not necessarily incompatible with the weak prescriptivist idea that good-sentences are often (or normally) used to commend (or recommend); it may even be compatible with the stronger thesis that "good" is (in its evaluative use, that is) essentially a commending word, i.e. that the evaluative use of language is essentially prescriptive. Or am I wrong in this, i.e. does prescriptivism entail that the conceptual gap between the evaluative and the normative is negligible after all?

[5]Cf. Haglund (1989), who writes: "[A] normative concept cannot be defined in purely evaluative (and/or descriptive) terms. You simply can't derive an 'ought' from an 'is' even if it is an evaluative 'is'" (p 107).

[6]However, this is not to deny that second-person-sentences like "this is good for you" are normally used prescriptively.

[7]This is closely connected to the ("externalist") idea that value-for-beliefs can not constitute reasons for action unless they are combined with some normative belief or other (or some desire or other; cf. below).

tant to notice that for such claims to be intelligible, they must be "based on" some malevolent directive, e.g., like "the worse off this person is, the better-period it is, the better the world is". Such malevolent directives may not be very plausible, but they are (and this is the point) both consistent and intelligible, e.g., someone may come to accept such a directive because he thinks that the person is evil and deserves to suffer.

But what if we (instead) have the "same-person case" in mind? Is it really "consistent and intelligible" for me to claim that someone has no reason whatsoever to perform a certain action that would benefit him, and harm no one else; or to claim that he has a reason to try to bring about a certain situation just because it is bad for him? Well, it is certainly odd, and it might be suspected that I don't really know what "good for" means, but as far as I can see, it is not entirely unintelligible[8].

(B) But to repeat, the fact that value-for-statements are not directives does not imply that they lack normative relevance. On the contrary, there are a number of possible reasons why we must regard a conception of (nonderivative) value-for as normatively relevant.

First, it supplies certain normative theories with the axiology they need, viz. theories that are (in a wide sense of the word) "teleological", i.e. theories that claim that moral rightness or practical rationality consist solely in the maximization of (and/or some distribution of) final value-for, like the self-interest theory or "interest-utilitarianism" (or, on the assumption that axiological individualism is true, utilitarianism).

And second, even if one does not accept any of these purely teleological theories, one can hardly deny that human welfare is something that an ethical theory needs to take into account: to make such a "purely deontological" claim is not merely implausible, there is also something inhuman (or "alien") about it. Regardless of what human welfare happens to consist in, it is simply something that an ethical

---

[8]This has certain implications for how the questions of prudential value should *not* be understood. For example, the question of prudential goodness should not be understood as a (normative) question about what it is rational to promote (or what we ought to promote), and the question of the good life must (in a similar way) be kept distinct from normative questions like "What kind of life is it most rational for a person to want?" or "What kind of life does a person have most reason to live?" (It is only if we accept the self-interest theory of rationality, i.e. if we assume that it is rational for a person to promote something if and only if it is good (best) for himself, that the two sets of questions "coincide").

theory must take into account. Or alternatively: There are a number of valid *prima facie* norms that involve a reference to human welfare, e.g., the idea that we ought to promote other people's welfare, especially the welfare of our own children, or that everyone has a reason to promote his or her own welfare. And in order to be of any practical use at all, these norms need to be combined with a conception of prudential value; this is the most important reason why such a conception must be regarded as normatively relevant. (We can also say that such a conception is normatively relevant "to the extent that" the most plausible ethical theory is a teleological theory).

To what extent, and in what ways, value-for-beliefs are action-guiding

The second claim on pp 399-400 can (like the first) be divided into two parts; (A) one negative claim, viz. the idea that a conception of value-for is not directly (and actually) action-guiding, and (B) one positive claim, viz. the idea that a conception of value-for is (in spite of this) "essentially action-guiding", and in a number of ways. Let us first look at the negative claim.

(A) We know that it is not necessarily implausible to regard the notion of value-period as a normative (or quasi-normative) notion, and value-period-statements as normative statements (or directives). Is it also the case that conceptions of value-period are directly action-guiding? Or more precisely: Is the belief that a certain situation X is good-period internally related to the desire that X obtains? Is value-period-internalism true?

If we modify what Thomson (1996) says about "Spirit-Emotivism" (on pp 113-118), we get three alternative formulations of value-period-internalism, namely

(1) Motivation Thesis: P's believing that X is good-period by itself motivates him to promote X;

(2) Wants Thesis: P's believing that X is good-period contains his wanting to promote X; and

(3) Weak Wants Thesis: There cannot be a person who has a fully fleshed-out conception of value-period like thine and mine but who

never wants to promote what he believes is good-period.

Now, in this context, it is not important to find out whether these three theses are true or false[9]. The point I want to make is (rather) that none of these three theses are implausible in any obvious way, and neither are emotivist theses like "value-period-sentences have no truth-values" or "a person who asserts a value-period-sentence merely displays an attitude, pro or con" (cf. Thomson (1996), pp 95-96)[10]. In short, as long as we have value-period in mind, it is *far from obvious* that views like emotivism and internalism are implausible[11].

Now, it is (I think) quite clear that these views are much more plausible than the views we get if we replace "value-period" with "value-for". For example, value-for-internalism is much less likely to be true than value-period-internalism. To see this, consider what value-for-internalism would be like[12].

There are two cases here that need to be distinguished, viz. the first-person case and the third-person case. Let us first consider what value-for-internalism would be like in the third-person case. Here are three alternative formulations:

(1) Motivation Thesis: P's believing that X is good for some other person Q by itself motivates him to promote X;

(2) Wants Thesis: P's believing that X is good for Q contains his wanting to promote X; and

(3) Weak Wants Thesis: There cannot be a person who has a fully fleshed-out conception of value-for like thine and mine but who never wants to promote what he believes is good for other people.

Now, (1) and (2) are obviously false. To believe that something is good for other people is not to have a motive for promoting it (and so on).

---

[9]Personally, I tend to accept (3) but reject (1) and (2); and here it doesn't really matter whether we add the assumption that P believes that X is realizable.

[10]The fact that all this holds in the normative case too may give some support to the idea that value-period is a normative notion (it may at least help explain why this idea is attractive).

[11]The same thing holds for prescriptivism: if we have value-period in mind, it is quite plausible to regard "good" as a commending word, and value-utterances as recommendations (or "commendations").

[12]Again, what follows is but a slight modification of what Thomson (1996) writes on pp 113-118.

For example, our believing that it is nonderivatively good for John to have pleasant experiences does not by itself motivate us to try to produce pleasant experiences in John. That is, it is quite possible to believe that a certain action would benefit a certain person, and harm no one else, and at the same time not be at all motivated to perform it. It is even possible to be motivated to promote a certain situation just *because* one believes that it is bad for a certain person (e.g., if one hates the person in question).

On my view, (3) is false too. A person who never wants to promote what he believes is good for other people is certainly not like you and me in all respects, but his conception of value-for may well be "just like thine and mine". So, if value-for-internalism is at all plausible, it is only in the first-person case.

So, what would value-for-internalism be like in the first-person case? Again, here are three alternative formulations:

(1) Motivation Thesis: P's believing that X is good for P by itself motivates him to promote X;

(2) Wants Thesis: P's believing that X is good for P contains his wanting to promote X; and

(3) Weak Wants Thesis: There cannot be a person who has a fully fleshed-out conception of value-for like thine and mine but who never wants to promote what he believes is good for himself.

Are these plausible views? Well, (3) seems plausible. Suppose that there is this person who never wants to promote what he believes is good for himself. This is a very odd person, and it is very likely that his notion of value-for is different from "thine and mine"; it is even possible that he does not understand what "value-for" means. (And if a child would consistently say that something is good for him, but that he does not want to have it, we would probably correct his language). But is his notion of value-for *necessarily* different from our notion? I don't think so. Think of a saint who has a strong desire to promote what he believes is good for other people, but who does not care about his own well-being. This saint may well have the same notion of value-for as we have (assuming that there was a time when he cared enough about his own well-being in order to learn the relevant concepts, that is).

However, I am prepared to accept the following "combined Weak

Wants Thesis": "There cannot be a person who has a fully fleshed-out conception of value-for like thine and mine but who never wants to promote what he believes is good for himself *nor* what he believes is good for other beings (e.g., other people)"[13].

(1) and (2) are much stronger claims than (3), and far less likely to be true. To see why (1) is implausible, consider the following question: Does John's believing that it is good for him to have pleasant experiences *by itself* motivate him to try to get more of these experiences? The answer is clearly "no". John's believing that something is in his own interest is likely to motivate him to try to promote it, but *only indirectly*, on condition that he cares about his own welfare, or desires to have a good life. (And not caring about one's own welfare is not a conceptual mistake, but a psychological disturbance; but again, we should think of what it takes to learn the relevant concepts). Or alternatively put: Value-for-beliefs are not action-guiding "in and by themselves"; to become action-guiding, they must be combined with some desire, normative belief, or the like.

And if internalism with regard to value-for is implausible, then it is also likely that certain other metaethical views become untenable if applied to value-for, e.g., views like emotivism or prescriptivism. That is, it does not seem very plausible to claim that value-for-claims have no truth-values, or that value-for-utterances are expressions of attitudes, pro or con (or that they are mere recommendations).

(B) But to deny that value-for-beliefs are *directly* action-guiding is not to deny that they are essentially action-guiding in a more indirect way. In fact, this is something that value-for-beliefs have in common with all other *kinds* of evaluative beliefs[14]. Now, that all kinds of evaluatives are action-guiding in some way, and to some extent, should not surprise us. After all, the only reason why evaluation is not a pointless activity is that we are (essentially) "decision-makers". If human beings weren't agents who had to make choices, there would probably be no such activity as evaluating. As an example, consider an evaluative statement of the form "X is a good hammer". To say that X is a good hammer is,

---

[13]And we may also add that it is (most certainly) impossible to have a fully fleshed-out conception of value-for if one has never (at any time) wanted to promote what one used to believe was good for oneself.

[14]That is, the claim is not that every *particular* evaluative belief is action-guiding.

supposedly, to say that it is good for use in hammering nails, i.e. that it is such "as to facilitate hammering nails in *well*" (Thomson (1996), p 134). This suggests that evaluations of hammers may well be action-guiding, but only if certain conditions are met, e.g., if there is someone who has this nail that he wants to hammer in. If no one had such wants, it would be entirely pointless (or even impossible) to go on evaluating hammers in this way. The same thing holds for value-for: The reason why it is not pointless to attribute value-for to things is that value-for-statements are often action-guiding. But they are (again) action-guiding only if certain conditions are met, e.g., if we care about people's well-being (and how could we not care about such a thing?[15]).

The different ways in which conceptions of prudential value are action-guiding, or: Why the questions of prudential value are important to us[16]

So, although value-for-beliefs are not directly action-guiding, it can hardly be denied that conceptions of prudential value are essentially action-guiding. This is the main reason why we regard the questions of prudential value as important questions, or alternatively put, this is the main reason why some of us regard it as important to find the most plausible conception of prudential value.

There are (I think) two major ways in which a conception of value-for may be (indirectly) action-guiding, viz. either by "hooking into" certain normative beliefs, or by "hooking into" certain desires. We will now take a closer look at *how* a conception of value-for may "hook into" our normative beliefs and/or desires. The most important ways in which a conception of prudential value may guide our choices are (I think) the following ones: (i) It guides our actions to the extent that we are self-interested; (ii) it guides our actions to the extent that we are benevolent; and (less importantly) (iii) it guides our actions to the extent that

---

[15]But do we really care about well-being *as such*? Isn't it rather the case that we all have some idea of what a person's well-being consists in, and that it is *this* that we care about? Well, this is often true, but it is certainly possible to be uncertain of what it consists in and still care about it, whatever it happens to consist in. And moreover, it is not likely that a change in a person's conception of well-being will have any major effect on how much this person cares about his own well-being (or the well-being of others).

[16]It is worth noting that in the following, the emphasis will be on why the central question (I) is important, i.e. I will not have very much to say about why (II) and (III) are important.

we are malevolent.

Now, I have already claimed that if we regard the question of prudential value as practically important, the main reason for this is that conceptions of prudential value are action-guiding. This means that there will be a correspondence between the three ways in which conceptions of prudential value are action-guiding and three kinds of reasons for regarding the question of prudential value as an important question. Let us now take a closer look at these three ways (three kinds of reasons).

(i) We all care about our own welfare: we are all, at least to some extent, *self-interested*. *Qua* self-interested (or *qua* egoists, if egoism is regarded as a matter of what one wants rather than what one does), we want to be as well off as possible. We also tend to believe that everyone has a prima facie reason to promote his or her own welfare, and some people (viz. the self-interest theorists) even think that maximizing one's own welfare is what practical rationality is all about, or (if they are also ethical egoists) that everyone "should" (has most reason to) act in such a way so as to maximize his own welfare.

But none of these self-interested desires and norms can guide our choices unless we have some idea of what it is that we (*qua* egoists) should want, promote, or maximize; *qua* self-interested, we need to know what our welfare consists in, and this is one way in which this question can become important to us[17].

(ii) Most of us are also, at least to some extent, *benevolent*: we care about other people's welfare, especially the welfare of the people to whom we are intimately related. Benevolence may take several forms, e.g., like love or altruism, and it may be paternalistic or not. To the extent that we *love*, or *care* about (and for) other people (like our partners, children, friends, or parents), we want them (on the traditional definition of love) to have good lives, and *qua* altruists, we give weight to the well-being of other people; we want them to be well off, and we also try (in our conduct) to benefit them. And if our benevolence is of a paternalist kind, we want to do what is good for other people, regardless of what they themselves think of this.

---

[17]It is not just the central question (I) that is important to a self-interested person, however. If he wants to be as well off as possible, he also needs an answer to (II), i.e. he needs to know how to determine just how valuable different possible outcomes are for him.

But it is not just that many of us are (to some extent) benevolent "in our desires", we are also more or less benevolent "in ethical theory". Most of us tend to think that everyone has a prima facie reason to promote other people's welfare, that it is morally wrong to harm other people, and that it is of special moral importance to promote the welfare of certain others, e.g., one's own children. Some people (e.g., certain kinds of interest-utilitarians) even think (roughly) that we are morally required to try to benefit everyone as much as we can (and to harm them as little as we can), or (if they are traditional utilitarians) that it is our moral duty to (try to) maximize the total amount of welfare in the world, regardless of whose welfare it is.

But again, none of these benevolent desires and norms can guide our choices unless we have some idea of what it is that we (*qua* benevolent benefactors) should promote or avoid, maximize or minimize; *qua* benevolent, we need to know what other people's welfare consists in, and this is the second way in which the question of the good life can become of practical importance to us[18].

(iii) To the extent that we are *malevolent*, we aim at what is bad for other people. If we are malevolent "in desire", we want certain others to be harmed (e.g., because we hate them, or because we are vengeful), and if we are malevolent "in ethical theory", we think we have a good moral reason to harm them, e.g., because they have done something that deserves some kind of punishment. Here, a conception of what is bad for us may come in handy; think of all the new little punishments that could be constructed!

So, these are (I think) the main reasons why the question of prudential value may be of importance *to us* (as we actually are). As they stand, these reasons do not tell us why we *should* regard this question as important. The fact that people regard a question as important for certain reasons does not in any way imply that these reasons are good reasons for regarding the question as important. This is rather obvious in the case of malevolence. The fact that people may regard the question of prudential value as important for malevolent reasons does not in any

[18]Or more specifically, the normative belief that we should promote other people's welfare makes the central question (I) important, and the belief that we should benefit other's as much as we can makes (II) important. There are also normative beliefs that make (III) important, however, e.g., the idea that we should spend more resources on the people who are worst off.

way mean that these reasons are good (or respectable) reasons for regarding the question as important. So we have to ask: Why *should* we regard the question of prudential value as important?

Why we should regard the questions of prudential value as important; normative relevance revisited

As I see it, the most important reason why it is important to find a plausible conception of well-being is that such a conception may be *"properly action-guiding"*, i.e. that it may help us to make better choices, e.g., with respect to one's own future life, the upbringing of one's children, or how to vote[19].

To say that a conception of prudential value is properly action-guiding is to say that it guides our choices by "hooking into" *valid* norms and/or *acceptable* desires (i.e. rational desires or morally acceptable desires). Suppose we have this idea that we ought to promote the welfare of one race and counteract the welfare of another race. In this case, a conception of welfare is not properly action-guiding, since the norm with which it is combined is not valid.

Now, to say that a conception of prudential value may guide our choices by "hooking into" *valid* norms is really the same thing as saying that it is normatively relevant. So the question arises: In virtue of what valid norms is a (plausible) conception of prudential value normatively relevant? What valid norms involve a reference to people's welfare?

As we have seen, there are three fundamental kinds of norms (*prima facie* or not) which involve an essential reference to human welfare; self-interested norms, benevolent norms, and malevolent norms. Are there valid norms of all three kinds? On my view, it is quite clear that there are valid norms both of the self-interested type and of the benevolent type (e.g., like "we ought to promote other people's welfare, especially the welfare of our own children" and "everyone has a prima facie reason to promote his or her own welfare"). Can malevolent directives also be valid? It is clear that malevolent desires can provide "subjective reasons" for acting, but it is far from obvious whether such desires can also provide objective reasons for acting (i.e. objectively valid direc-

---

[19]However, it is hardly *necessary* to have a plausible conception of prudential value in order to make good decisions. And it is (of course) far from *sufficient*; in order to make good choices, there are *so* many other things we have to know (besides what has final value for us).

tives). Do we ever have a good moral reason to harm someone (regardless of whether this makes other people better off)? Frankly, I don't really know[20].

A comparison with Scanlon's perspectives

At this point, it might be informative to compare the different ways in which the question of prudential value may be of normative relevance with what Scanlon (1993) has to say about the different "points of view" from which the question of the good life can be asked[21]. After having pointed out that "there are a number of different standpoints from which the question of what makes a person's life better /.../ might be asked" (p 185), he gives us five examples of such standpoints (or alternatively: normative contexts). Scanlon writes:

> It [the question of what makes a person's life good for this person] might be asked from the point of view of *that person herself*, who is trying to decide how to live. It might be asked from the point of view of *a benevolent third party*, a friend or parent, who wants to make the person's life better. It might be asked, in a more general sense, from the point of view of *a conscientious administrator*, whose duty it is to act in the interest of some group of people. It might be asked, again in this more general sense, by *a conscientious voter* who is trying to decide which policy to vote for and defend in public debate and wants to support the policy which will improve the quality of life in her society. Finally, the question of what makes a person's life better also arises *in the course of moral argument* about what our duties and obligations are, since these duties and obligations are surely determined, at least to some extent, by what is needed to make people's lives better or, at least, to prevent them from being made worse (ibid., p 185, my italics).

---

[20]At this point, it is worth pointing out that there are also other reasons why we should regard it as important to find a plausible conception of well-being. For example, a conception of the good life may also be legitimately used to evaluate and criticize circumstances, or in particular, to criticize societies (or cultures). This kind of *social (or cultural) criticism* would be of the following form: "Society S (or culture C) is bad (should be changed in a certain way) because it (e.g., its institutions) makes it difficult or impossible for people (primarily for the citizens in S, but also for outsiders) to lead good lives, or good enough lives".

[21]In this context, it is of little or no importance to distinguish between the question of prudential value and the question of the good life.

If we compare this with the different ways in which the question of the good life may be of normative relevance, we see that Scanlon covers more ground than I do. The two "general senses" in which the question might be asked (i.e. from the point of view of a conscientious administrator, and by a conscientious voter) have no obvious counterparts on my list - my closest point is "the utilitarian angle"- but the other three points of view are more or less identical with items on my list. First, to say that the question might be asked from the point of view of that person herself is (roughly) to say that we might ask it *qua* self-interested; second, to say that it might be asked from the point of view of a benevolent third party is (roughly) to say that we might ask it *qua* loving and caring, or *qua* altruists; and third, to say that the question might arise in the course of moral argument about what our duties and obligations are is (roughly) to say that we might also ask it *qua* moral theorists (e.g., *qua* utilitarians)[22].

So, does this mean that I have to supplement my list by adding that the question of the good life may also be of *political relevance*, in either of the two ways suggested by Scanlon? I think not. As Scanlon himself points out (again on p 185), there are several ways in which the question of the good life may be understood, e.g., it may not just be interpreted as "What makes a life a good one for the person who lives it?", but also as "What kinds of circumstances provide good conditions under which to live?". Now, my suggestion is that when the question of the good life is asked by conscientious administrator or a conscientious voter, then it must be understood in the second of these two senses. The reason for this is that from a (concrete) political perspective, it makes much more sense (and it is much more fruitful) to ask what kinds of circumstances that provide good conditions under which to live than to ask what it is that makes a life have final value for the person who lives it.

To sum up: We have now seen that there is a number of ways in which the questions of prudential value are of normative relevance: They are important to us *qua* self-interested, *qua* benevolent third parties (paternalists or not), and *qua* moral theorists (utilitarians or not), and they *may* also be of some importance to us *qua* administrators and

[22]However, we should not forget that the difference between a moral argument and an argument between moral theorists might be considerable.

qua voters.

Is there such a thing as the most plausible answer to the questions of prudential value?

This gives rise to the possibility that there is no such thing as *the* most plausible conception of prudential value (common to all perspectives), i.e. that the most plausible answer to the questions of prudential value may (instead) vary with the point of view from which these questions are being asked. Scanlon warns us of this possibility when he writes that "the plausibility of various answers can be strongly influenced by the point of view of the question[23], and unnoticed shifts in point of view can drive us back and forth between different answers" (p 185).

So we have to ask ourselves whether there is such a thing as *the* most plausible answer to the questions of prudential value[24]. Or formulated in terms of the different ways in which a conception of well-being may be normatively relevant: Is the welfare that we have a reason to promote *qua* self-interested *the same human welfare* that we have a reason to promote *qua* benevolent third parties? And are these "welfares" *the same welfare* that the utilitarian thinks we ought to maximize? And so on.

These are very difficult questions, however, and I will not try to answer them in this book. Suffice it to say that *if* Scanlon's suspicion is well-founded, i.e. *if* "the plausibility of various answers" is (in fact) "strongly influenced by the point of view of the question", then this implies that the question "What has final value for a person?" must be understood in relation to some normative theory (or *prima facie* norm) N, i.e. it must be interpreted as "What conception of prudential value makes (if embedded in N) N most plausible?". This seems to imply that arguments for and against different conceptions of well-being must be (to a considerable extent) normative rather that straight-forwardly evaluative. I tend to believe that Scanlon's suspicion is not well-founded, however, at least not as long as we remain within the intra-personal domain[25].

---

[23]Or more precisely, there may not be such a thing as *the* question, viz. because different points of view give rise to different interpretations of the "question", i.e. to different questions.

[24]Or whether it is appropriate to talk of *the* questions of prudential value. Cf. note 23.

[25]It is possible that the theory of welfare which is most "normatively adequate" in an intrapersonal context differs from the theory that is most plausible in an

The idea that value-for is fundamentally comparative

Let us now turn to the idea that goodness and badness are "fundamentally comparative". How should this idea be understood? Well, the idea is that goodness and badness can be defined in terms of betterness, or that there is (so to speak) nothing more to goodness than betterness[26]. Is this a plausible claim? Well, not if it is understood as a claim about all kinds of goodness (and badness). As Thomson (1994) points out, a good K is not (by definition) a K that is better than most Ks, e.g., it could well have been the case that all hammers are good hammers (p 12). But the fact that the claim is not valid for all kinds of goodness does not imply that there are no kinds of goodness for which it is valid. So the question arises: Are there kinds of goodness for which the claim is valid? In particular, is the claim that there is no more to goodness-for than betterness-for a plausible claim?

Here, we have to ask ourselves what philosophers like Chisholm and Sosa (1966) really have in mind when they claim that betterness (or more specifically, betterness-period) is fundamental to goodness (and badness). What is it that they say about goodness-period and betterness-period when they make their claim? Well, the idea seems to be that if we knew, for all possible situations, the structure of the betterness relation between them (if we knew, for every pair of situations, whether one was better than the other), then we would know everything there is to know about the goodness of the situations. In short, "[a]nything that can be said in terms of the one-place property 'good' can be said in terms of the two-place betterness relation 'as least

---

interpersonal context. For example, it seems that the version of the actual desire theory (see section 1.2) that is most plausible in an intrapersonal normative context may not fully coincide with the version that is most plausible in an interpersonal context (cf. section 5.1.2).

[26]That is, the claim is *not* that the question of what it is that makes one thing better (in way W) than another thing can (or must) be answered before the question of what is good (in way W), or what it is that makes a thing good (in way W). On the contrary, it seems (as we will see below) that it is the other way around, e.g., it seems (in the case of prudential value) that (I) the question of what is good for us must be answered before (II) the question of how we should determine whether a certain situation is better for a certain person than another situation. This might *seem* inconsistent with the idea that betterness is primary to goodness, but it is not.

as good as'"[27]. But does this really agree with the fact that goodness-period-statements contain something that most betterness-period-statements do not contain, viz. an implicit reference to a zero point (if we are told that a certain situation is good, then we are also told (implicitly) that it is better than nothing)? I think it does. Or more specifically, I believe that Chisholm's and Sosa (1966) have shown how goodness-period and badness-period can be defined in terms of betterness-period, viz. in the following way: We first define "indifference" in terms of "betterness": A state of affairs X is indifferent if its realization has the same value as its non-realization (or better: if its presence has the same value as its absence), i.e. if X is not better than not-X, and if not-X is not better than X[28]. "Good" and "bad" can then be defined in terms of "better" and "indifferent": "Let us say that a state of affairs is good provided it is better than some state of affairs that is indifferent, and let us say that a state of affairs is bad provided that some state of affairs that is indifferent is better than it" (p 246).

Now, this seems to hold for value-for as well. That is, it seems that goodness-for is (like goodness-period) fundamentally comparative, i.e. there is (in the above sense) nothing more to goodness-for than betterness-for. This may throw some light on (I), but apart from this, it has really no relevance for how the central questions of prudential value should be understood. For example, we can not conclude that (II) is more fundamental than (I).

There are reasons to believe that *particular* comparative evaluations are of more practical importance than *particular* non-comparative evaluations, however, viz. because the former are normally more action-guiding (and more normatively relevant) than the latter. For example, if someone has to make a choice between two possible situations, what he needs to know is not whether these situations are good-period or bad-period, but which one is better-period (e.g., it is, in a case like this, not very informative to be told that they are both good). This seems to hold for value-for as well: it seems plausible to assume that particular betterness-for-claims are more normatively relevant than particular

---

[27]The formulation is from Broome's *The Value of Living* (a manuscript written in 1993), p 30.

[28]On the assumption that the value of X and the value of not-X are comparable, that is.

goodness-for-claims[29]. Or in the special case of well-being, it is normally more normatively important to determine changes in well-being than to determine levels of well-being. This does not mean that we can do without non-comparative evaluations in ethics, however. For example, it is sometimes of normative importance that we can determine whether a certain life is "good enough" or not, or whether it is "worth living" or not. The notion of "a life worth living" is highly relevant in the debate on euthanasia, and the notion of "a good enough life" is (I think) made relevant by some normative view like "Only people whose lives are not good enough are entitled to certain kinds of support from society" or "If a person has a good enough life, he does not have a strong reason to try to improve his own life further; instead, he should use his energy to promote the welfare of others who are less fortunate than he is". (If no such normative "principle" is valid, the question of the good enough life does not really belong in ethics).

This ends our discussion of the relations between goodness and betterness. Let us now turn to the idea that value-for is supervenient in a certain way.

Value-for is supervenient, but in a very special way

It is often pointed out that no matter what kind of value we have in mind, the value of a particular valuable thing always *supervenes on* the natural (or descriptive) features of this thing. In the case of goodness, this means that all good things are good in virtue of certain features that they have. And since we can say that the features in virtue of which something is good *make it good*, it is natural to refer to these features as *good-making characteristics*. (This is the reason why it is always legitimate to ask what it is that makes a good thing good).

To say that a certain thing is good in virtue of certain features that it has is to say that it is good *because* it has these features, and this means that the idea of supervenience can also be formulated in terms of reasons: If something is good, then it is good for a reason, and this reason is of the form "it has the natural features $F_1 \ldots F_n$". This suggests

[29]But again, this does not imply that general betterness-for-claims (e.g., answers to (II)) are (in any way) more fundamental than general goodness-for-claims (e.g., answers to (I)). The reason for this is that it can (in this context) be plausibly assumed that the more there is of a good thing, the better. Cf. the section on supervenience below).

that reasons for particular evaluations are normally descriptive. A good term for the facts that corresponds to such descriptive statements (e.g., the facts that corresponds to "X has F_1", "X has F_2", and so on) is *"good-making fact"*.

The reason why descriptive sentences can count as reasons for evaluative sentences is that there are so called *"standards of goodness (or badness)"*. Such a standard tells us that certain features (qualities or relational properties) should count as good-making (etc.), and in this way, the standard can be seen as "constituting" some kind of "logical" connection between the descriptive reason and the evaluation (but cf. note 30). Standards of goodness are of many different kinds, and what features that are counted as good-making depends on what kind of goodness we have in mind, and in the case of attributive goodness, what kind of thing the object is (what the class of comparison is).

Depending on how they "connect" (particular) evaluations with their reasons, standards of goodness can be of different kinds. To illustrate this, let us restrict our attention to attributive goodness, i.e. to evaluations of the form "X is a good K". Now, as far as I can see, the connection between such an evaluation and its reason (e.g., "X has $F_1 \ldots F_n$") may be of three different kinds:

(i) For X to be a good K, it is *necessary* that it has $F_1 \ldots F_n$ (or some combination of them). In this case, the standard of goodness (which is common to all Ks) is of the form "The presence of feature F in a K is necessary for it being a good K".

(ii) That a K has $F_1 \ldots F_n$ may also be seen as a *sufficient condition* of it being a good K. In this case, the standard of goodness is of the form "The presence of feature F in a K is sufficient for it being a good K"[30].

(iii) But the fact that "X has $F_1 \ldots F_n$" is a reason for "X is a good K" does *not* imply that the presence of $F_1 \ldots F_n$ in X is either necessary or sufficient for X's being a good K. There is at least one more way in which a natural feature of X can make (or help make) X a good K, viz. by being what Beardsley (1981) calls a *merit*. A meritorious feature of something contributes to its goodness, but not by being a necessary or sufficient condition for goodness (and in a similar way, a defect, or defective feature, of a thing is a feature that detracts from its goodness,

[30]It is only when a standard of goodness takes this form that it can connect a particular evaluation and a reason in a strictly logical (i.e. deductive) way.

or contributes to its badness). That is, a principle about merits and defects (or what Beardsley, in an aesthetic context, calls a *Canon*) for Ks does not give us any necessary or sufficient conditions for goodness or badness of Ks. This is how Beardsley himself puts it (for K = poetry):

> Thus if there is a general principle - stated as, "Vague themes are always defects in poetry," or, "Grand imagery is always a merit in poetry" - this principle does not mean that these features are either *necessary* or *sufficient* conditions of goodness in poetry, but only that, other things being equal, their presence makes it better or worse (pp 464-465)[31].

What has been said this far can be regarded as "typical supervenience-talk": (a) It tends to focus on goodness (and badness) rather than betterness, (b) goodness is typically conceived of as a property rather than as desirability (or to-be-pursuedness), and it is typically attributed to things rather than to situations (which explains why it tends to focus solely on attributive goodness). So, let us now see what happens to the idea of supervenience if we widen our scope a little.

(a) To say that the betterness relation is a supervenient relation seems to make perfectly good sense; it simply means that we can always give reasons for our particular comparative evaluative statements: if a particular thing X is better than a particular thing Y, then it is better for a reason. But when we ask ourselves what kind of reason, matters becomes tricky. Here, we may not be able to say anything about goodness in general: we probably have to consider different kinds of goodness separately. So, what if we have "attributive betterness" in mind? What does a reason for "X is a better K than Y is" look like? Well, we can hardly refer to better-making features, or to the relation between X and Y. My guess is that we have to appeal to good- and bad-making characteristics in this case too, i.e. if X is a better K than Y, the reason for this is that X has (so to speak) "more of" the good-making features for Ks, and/or "less of" the bad-making features for Ks. That is, it seems that in the case of "attributive value", giving reasons for non-

[31]He also adds that it "does not seem that the contribution of each feature of an aesthetic object can be considered in an atomic fashion". This seems to suggest that a certain feature can be a merit in some Ks and a defect in other Ks, but it is doubtful whether such a radical "evaluative holism" (or "particularism") is consistent with the idea that other things being equal, the presence of a merit in a thing makes this thing better.

comparative particular value judgements is (somehow) primary to giving reasons for comparative particular value judgements (cf. notes 26 and 29). (We will soon try to find out whether this holds for value-period and value-for as well).

(b) Even though the idea of supervenience is often "based on" the conception of value as property, it seems that is also compatible with the conception of value as desirability (or to-be-pursuedness). For example, to say that value *qua* to-be-pursuedness is supervenient is simply to say that if we have a reason to promote (e.g., to try to maintain) a certain particular situation X[32], then there is a reason for this, and the reason is that X has certain features (intrinsic or relational)[33]. But what is it to be a feature of a particular, concrete situation? Well, in my view, a feature of a particular situation is simply a correct description of it, or it corresponds to such a description, e.g., if a certain situation can be correctly described as "Bert has a pleasant experience", then the corresponding fact is a feature of the situation[34].

In a similar way, we can regard betterness-period as supervenient too, i.e. we can safely assume that if a particular situation X is better-period than another particular situation Y, then there is a reason for this. But what kind of reason? My guess is that we have to appeal to good- and bad-making characteristics in this case too: if X is better-period than Y, the reason for this is that X "contains" more good, or

---

[32]It is worth emphasizing that in this context, it is only *particular* situations that are of interest, and for two reasons: First, the carriers of desirability (or to-be-pursuedness) are situations. Second, the idea of supervenience only makes good sense if it is conceived of as an idea about the goodness of particular, concrete things. That is, the plausible claim that is hidden in the idea of supervenience is that we can always give reasons for our particular evaluations; the claim is *not* (or should not be) that we can always give reasons for our general evaluations, or standards of goodness. Now, this is (of course) not to say that it is impossible to give reasons for standards of goodness (i.e. for general evaluations of the form "all situations of type X have positive final value-period"), but rather: The question of whether it is possible or not to give reasons for our standards of goodness has nothing to do with the issue of supervenience, and it is (moreover) likely that these reasons will be very different from reasons for particular evaluations.

[33]This strongly suggests that it is not just attributive goodness that is supervenient: Predicative goodness can (and should) be regarded as supervenient too, i.e. if some particular situation is good in the predicative sense, then we can safely assume it is good for a reason, viz. because it has certain features.

[34]This would explain why it doesn't make any sense to talk about *features of situation-types*: no clear distinction can be made between situation-types and situation-descriptions. It also gives rise to the suspicion that a concrete situation is not really as distinct from its (correct) descriptions as one might believe.

less bad, than Y. This suggests that we can make the following *general* claim about giving reasons for particular evaluations: *Giving reasons for non-comparative particular value judgements is often (perhaps always) primary to giving reasons for comparative particular value judgements*[35].

Let us now turn to the supervenience of goodness-for (badness-for) and betterness-for. Like all other kinds of goodness, *goodness-for* is supervenient, but only in "particular cases". That is, if a certain particular situation is good for a certain person, then it is (supposedly) good for the person in virtue of certain features that it has. Or in terms of reasons: If X is good for P, then it is good-for-P for a reason, and this reason is of the form "X has the natural features $F_1...F_n$".

But this is only partially true. As we have already seen (e.g., on p 3), goodness-for should be conceived of as a *relation* between an object and a subject, and this means that *the features of P* must also be taken into account. So, it is true that if a particular situation X is good for a particular person P, then it is good-for-P for a reason, but this reason is not of the form "X has the natural features $F_1...F_n$", but (rather) of the form "X has the natural features $F_1...F_n$, *and P has the natural features* $G_1...G_m$"[36]. This is how the supervenience of goodness-for differs from the way (or ways) in which other kinds of goodness are supervenient: Whether X has value for P does not just depend on what X is like, but also on what P is like[37].

Now, this suggests that we can also give a certain type of reason for more general value-for-claims, viz. we can say that all situations of a certain type are good for a particular person P in virtue of P's features,

---

[35]The "ultimate" reason for this is (probably) that we tend to think in terms of goodness and badness rather than in terms of betterness. In particular, we tend to understand "X is a better situation than Y" as "X has more good features (contains more good things) and/or less bad features", or alternatively put (as Björn Haglund has pointed out), we tend to believe that the betterness-relation is intrinsic (rather than extrinsic) in the following sense: It supervenes on the features of the objects related, and not on any (natural) relations between the objects.

[36]Where the features of X may be relational, but hardly (I think) the features of P.

[37]This does not mean that other kinds of "goodness-supervenience" are *causally independent* of what we are like, however. For example, the reason why sweetness, juiciness, etc. are regarded as merits in strawberries is that we happen to have the kind of taste buds we have. It is not unlikely that this goes for all standards of goodness (no matter what kind of goodness we have in mind): the fact that we count certain features as good-making and other features as bad-making can probably be *explained* in terms of what we are like.

and we can say that all situations of a certain type are good for all of us in virtue of our common human nature (It should be noticed that this does not *follow* from the idea of supervenience, however). At this point, I would like to make an even stronger claim, however: No "semi-general" value-for-statement of the form "all facts of type X have final value for a particular person P" can be sufficiently justified unless the justification is "subject-oriented", i.e. unless it takes the form "all facts of type X are good for P *because* P has the natural features $G_1 \ldots G_m$"[38]. And the same thing holds for fully general value-for-statements of the form "all facts of type X have final value for all human beings" (i.e. for all answers to the central question (I)): No such claim can be completely justified unless it takes the form "all facts of type X are good for all human beings *because* human beings are constituted in such-and-such a way"[39]. (We will return to this central justificatory principle on a number of occasions).

Now, assume that the general claim about giving reasons for particular evaluations (on p 420 above) is true, i.e. that reasons for comparative particular value judgements are (so to speak) "parasitic upon" reasons for non-comparative particular value judgements. In the case of value-for, this would mean that all standards of prudential betterness can be derived from standards of prudential goodness, *plus* the idea that if X is nonderivatively good for P, then the more there is of X, the nonderivatively better for P. That is, in order to find an answer to the comparative question (II) "How do we determine just how valuable a certain possible situation is for a certain person?", we must first find an answer to the noncomparative question (I) "What is nonderivatively good and bad for a person?" (cf. notes 26, 29, and 35 above).

---

[38]Assuming that X does not contain any essential reference to P's intrinsic features, that is. Cf. note 39.

[39]This seems to presuppose that X does not contain any reference to our intrinsic features. For example, it is not easy to see how claims like "It is good for all of us to have our desires fulfilled (or to have our basic needs satisfied)" can be justified in terms of our intrinsic features. In fact, it is hard to see how they can be justified at all.

Appendix C
Well-being, value-for, and time

In chapter 1, it was assumed that the comparative question of well-being (i.e. (III) "How do we determine just how valuable a person's existence is for this person?") can be interpreted in three different ways, viz. as (i) "How do we determine just how well off a certain person is at a certain time?", or as (ii) "How do we determine just how valuable a person's existence (life) over a certain period of time is for this person?" (a special case of which is (iii) "How do we determine just how valuable a certain life as a whole (from birth to death) is for the person who lives it?").

Now, it may seem that these three interpretations all make sense, but is this really so? For example, it seems intelligible to attribute aesthetic value or value-period to whole lives, but does it really make sense to attribute final value-for to lives as temporally extended wholes? Or more specifically, does it makes sense to talk about (i) the final value-of-a-life-at-a-certain-time-for-the-person-who-is-living-it, and/or (ii) the final value-of-a-life-over-time-for-the-person-who-is-living-it, and/or (iii) the final value-of-a-life-as-a-whole-for-the-person-who-is-living-it? To be able to answer these questions, we need to make some reflections on value-for and time.

When we talk about the value of a life for the person who is living it, it is likely that the term "life" refers to the person's particular, concrete life (but it may also, at times, refer to some possible future life), and that the term "person" refers to some particular, concrete person. That is, it seems that the value-for-statements that are of interest in this context are either of the form "the actual particular situation X is (or was) good for the particular person P" or "the possible (hypothetical) particular situation X would (or will) be good for the particular person P".

I now want to suggest that *statements of these forms can only be intelligible and/or plausible on condition that the temporal relation between X and P meets certain requirements.*

Let us first establish that X and P both have (in this context) "temporal location", i.e. that they are the kinds of entities that stand in temporal relations to other entities. First, all actual and hypothetical situations can be thought of as obtaining at some time or other, i.e. we

either have to think of X as X-at-t or as X-between-t_1-and-t_2[1]. Second, if P is a particular person, then we must (in this context) think of P as temporally located too, viz. as P-at-t, where t is either a particular time or a variable bound by a quantifier (as in "all situations of type X are good for P at any t")[2]. The reason for this is that when we say that something is good for a particular person P, there is always an implicit reference to time in the case of P too, i.e. we can (and should) think of all goodness-for-P as goodness-for-P-at-t. This means that if a goodness-for-statement is fully spelled out, it is of the form "X-between-t_1-and-t_2 is (was, or will be) good for P-at-t_3", where t_1, t_2 and t_3 may (but need not) be identical.

It is important to see that the idea that the logical subject of goodness-for-people-statements are persons-at-certain-times does not have any ontological implications. In particular, it seems compatible with the idea that persons are atemporal objects which persist over time, i.e. it does not force me to accept the idea that persons are temporally extended entities which can be divided into small temporal "slices". (As I see it, all that "P-at-t" refers to is P as he or she is at t).

So, what are my reasons for regarding P-at-t as the logical subject of value-for-statements? Well, my argument can be divided into two steps. First, it makes perfectly good sense to say that a certain situation is good for P at one time, but not at another time (e.g., because P has changed in the meantime); and it seems plausible to say that it cannot have final value for me now that something happened in the past, or that it will happen in the future. This suggests that the complete analysis of goodness-for-statements has to include another time reference (besides the time at which X occurs). There are two options here: either we regard P as temporally located (i.e. we think of all goodness-for-P as goodness-for-P-at-t), or we regard the value as temporally located (we think of all goodness-for-P as goodness-at-t-for-P). The second step consists in trying to show that we should prefer the former alternative to the latter, viz. for the following (simple) reason. If value-for is

---

[1]Where the fact to which "X" refers does not contain any reference to time. It may well contain a reference to P, though.

[2]However, we can not (in this context, that is) think of a particular person P as extended in time, i.e. we can not think of him as P-between-t_1-and-t_2. It is also worth adding that if X-at-t is good for P-at-t, this does not necessarily mean that the two occurrences of "t" refer to the same thing: While the "t" in "P-at-t" refers to a point in time, the "t" in "X-at-t" may refer to a short period of time.

a relation, it seems more attractive to think of the relata as temporally localized than to think of the relation itself as temporally localized: in particular, isn't it rather odd to regard the relation as temporally localized if one of the relata is not temporally localized? (This is the main reason for regarding all goodness-for-P as goodness-for-P-at-t. It is not the only reason, however, e.g., it also seems that the idea has a certain "explanatory power").

Now, if we look at statements of the form "X-at-t_1 is (was, or will be) good for P-at-t_2", we can see that there are three possible temporal relations between X and P, viz. (i) $t_1 = t_2$ (X and P are simultaneous), (ii) t_1 is earlier than t_2 (X is in P's past, or before P's birth), and (iii) t_1 is later than t_2 (X is in P's future, or after P's death). If we also consider statements of the form "X-between-t_1-and-t_2 is (was, or will be) good for P-at-t_3, we can see that there is a fourth possibility, viz. (iv) t_3 is later than t_1 but earlier than t_2 (i.e. P is, so to speak, in the midst of X)[3].

Let us now consider these possibilities one at the time, to see whether they make sense (and whether they are plausible):

(i) It obviously makes sense to say that it is (was, or will be) good for P-at-t that X obtains (obtained, or will obtain) at t, e.g., it makes perfectly good sense to say that it is good for me-now that I have a pleasant experience right now.

(ii) Does it also make sense to say that it is good for P-now that some situation X occurred in P's past? It is clear that past situations (e.g., a happy childhood) can affect P's present situation favourably, and that they can (for this reason) have instrumental value for P-now. However, it is not likely that past situations can have final value for P-now. Or am I mistaken in this? Think of an old woman who looks back on her life and regards it as good. What kind of goodness is it likely that she attributes to her past life? Well, it is not entirely implausible to assume that it is (at least in part) goodness-for, and that it is (moreover) not merely instrumental goodness-for. This suggests that it might make sense to say that it has final value for P-now to have lived the life she has. However, it is also possible that evaluations of this kind are all based on some "mistake". This is the view I tend to adopt.

---

[3]There is (of course) also a fifth and a sixth possibility, viz. (v) $t_3 = t_1$ (P is, so to speak, in the beginning of X), and (vi) $t_3 = t_2$ (P is at the end of X). I don't think these possibilities are interesting enough to merit their own discussions, however.

(iii) So, what about future situations? Does it make sense to say that it is better for me-now if I have 40 years left to live than if I have 20 years left to live, or that the shorter my future life is, the worse it is for me-now? Does it make sense to say that it is bad for a person to die (i.e. that his future life is limited), or that it is bad for me-now that I will loose my job next year? Well, it may seem so[4], but this is a tricky matter, and for the following reason. Suppose that the longer my future life is, the better it is for me-now. We then have to ask "in what way" it is supposed to better for me-now. It is hardly nonderivatively better for me-now. So, is it (then) derivatively better for me-now? Well, a longer life makes more good things possible, and it may therefore seem that it is derivatively better for me to live longer. But is it really derivatively good for me-*now* that I-*then* will have a good life then? I think not.

(iv) Does it make sense to claim that it is better for me-now if the pleasure I experience right now lasts for four hours than if it lasts for two hours; or that it is better for a person if he lives for 80 years than if he lives for 70 years; or that it is better for a person if his life as a whole is organically unified than if it is not? And are these plausible claims? I think not, not if the person is (so to speak) in the midst of the good (better, or worse) situation. What I have in mind is this: It may well be better for me if I live until I'm 80 than if I live until I'm 70, but if it is, this is solely due to the fact that I have 10 years longer left to live in the former case. In short, it is only the future that is of interest to me-now. Or alternatively, it is better for a person to be 60 and live for 10 more years than to be 75 and live for 5 more years. On my view, this suggests that it is implausible (it may even be nonsensical) to talk about the value-for of a life as a whole, especially if we have final value-for in mind. If a person is in the midst of his life, then his life-as-a-whole is partly "in the past", partly present, and partly "in the future", and this is why it doesn't make sense to talk about the value of such a strange entity *for the present person*[5].

---

[4]As Sven Danielsson once pointed out (in conversation), it seems that the person who has 10 years left to live is *in a better position* than the person who has only 5 years left to live.

[5]But as Björn Haglund has suggested (in conversation), it would perhaps make sense to attribute value-for to a life-as-a-whole if the whole life is a future life, e.g., when one has (*qua* potential parent) a choice to "create" several different possible lives-as-wholes, or if a person is about to be reborn and has a choice between different possible lives-as-wholes.

To conclude, if a person P is in the "midst" of his life, it does not seem to make any sense to attribute value-for-P (derivative or nonderivative) to his life-as-a-whole. However, it clearly makes sense to attribute value-for (derivative as well as nonderivative) to a-life-at-a-certain-time, e.g., to say that the life that P lives at a certain time has final value for P-at-that-time. Does it also make sense to attribute value-for to a life-over-time? Well, it seems intelligible to attribute derivative value-for to past lives, but I tend to reject the idea that past lives can have final value for present persons. It *may* also make sense to attribute derivative value-for to future lives (e.g., it might make sense to say that it would be derivatively good for P-now if his life would be good over the next five years), but it is surely absurd to attribute nonderivative value-for to future lives. In short, the only life that can have final value for P-at-*t* is the life P lives at *t*. A person's past life can only have derivative value, and the same thing holds (at best) for a person's future life[6]. Which suggests that the comparative question of well-being (III) should be understood as "How do we determine just how well off a certain person is (on the whole) at a certain time?".

It is worth noting that the idea that all goodness-for-P as goodness-for-P-at-*t and* that X cannot be good for P-at-*t* unless it occurs at *t* has several implications (some of which may be hard to swallow). Here are two examples:

(1) As we have already seen, it doesn't seem to make sense to attribute final value-for-P to temporally extended situations, e.g., to lives over time. Or alternatively put, all well-being is well-being-at-*t*. So, what if we simply add P's well-being-values at different times: doesn't this give us a measure of P's well-being "over time"? Of course we can obtain a numerical sum in this way, but since there is no such thing as a temporal super-subject P-over-time, this sum does not reflect anything

[6]That is, X can only have final value for P-at-*t* if X obtains at *t*. Now, as I see it, this is compatible with the idea that the value of X-at-*t* for P-at-*t* may well be dependent on how X is viewed, or described. For example, a certain concrete situation may well be more valuable for a certain person if it is regarded as a part of a larger whole (e.g., as part of a project, or a "life plan") than if it is regarded in isolation, and the value of a situation may also depend on things like how it has originated, or whether it is regarded as a "new beginning". In short, I think my "present-oriented" theory of final value-for can account for the fact that we are future-oriented as well as "past-oriented", and that this is manifested in (valid) evaluations.

real[7]. (This does not mean that the sum lacks normative relevance, however).

(2) We cannot say that death is not an evil for the person who dies, nor that it is better for a person to have a longer than a shorter life. Now, this is not to deny (a) that it is better-period that a person lives longer, and that this "value" is (in some sense) localized to (or possessed by) the person, nor (b) that it is rational for a person to try to stay alive, i.e. to avoid his own death (That is, value-for should not be conceived of as that which it is rational for an egoist to maximize; this would be to "normativize" the notion of value-for in an unduly manner)[8].

---

[7]Cf. the analogy with temperature. It makes perfectly good sense to say that at each point in time, this room has a certain temperature (which can be measured on an interval scale). There is no such thing as temperature-over-time, however (if we add the measures at different times, this sum will not reflect anything real; it is possible, however, to use the measures to calculate the average temperature, and so forth). I owe this observation to Björn Haglund.

[8]Other implications are: It cannot have nonderivative value for a person to have his prospective (or retrospective) desires fulfilled; what happens after a person's death cannot have value for this person (even if it happens to be the object of his last wish); and it is not nonderivatively worse for a person to suffer for a longer time than for a shorter time.

Appendix D
Subjectivism and Objectivism
Some other important questions of well-being

It has already been pointed out (in chapter 1) that some of the central questions in the recent philosophical literature on well-being are not included in the central questions of this book. The main purpose of this appendix is to make these questions explicit, mainly to get a better understanding of what needs to be included in a complete conception of well-being (prudential value).

Most of the modern discussion of prudential value (or "well-being") is (it seems) based on Parfit's (1984) distinction between three kinds of conceptions of the good life (or, as he prefers to call them: "theories of self-interest"), viz. Hedonistic Theories, Desire-Fulfilment Theories, and Objective List Theories. This is how Parfit characterizes these theories:

> On *Hedonistic Theories*, what would be best for someone is what would make his life happiest. On *Desire-Fulfilment Theories*, what would be best for someone is what /.../ would best fulfil his desires. On *Objective List Theories*, certain things are good or bad for us, whether or not we want to have the good things, or to avoid the bad things (p 493).

Of these three kinds of theory, it is the hedonistic theory that is most easy to understand, and the reason for this is first, that it is a substantive evaluative theory, and second, that it is nothing but a substantive evaluative theory. That is, it gives us straight-forward answers to the substantive evaluative questions of prudential value, i.e. to the central questions (I)-(III), and it does not purport to answer any other questions, least of all any questions of a metaethical nature.

The other two theories can (obviously) *not* be regarded in this way. It is true that every Objective List Theory makes substantive claims about what is nonderivatively good and bad for a person (in this respect, it does not differ from hedonism), but this is not all there is to it. Above all, the fact that all objective list theories are substantive evaluative theories is not what *makes* them objective, and this suggests that the substantive evaluative questions are not the only questions that these theories purport to answer.

If we look at the Desire-Fulfilment Theory, we see that it differs even

more from hedonism. First, it does not make any *substantive* evaluative claims at all[1], and second, it is (like the objective list theory) more than just an evaluative theory; it also makes other kinds of claims about the connection between value-for and desire (claims which may well be of a metaethical nature), and it is in this area that the most central differences between the desire theories and the objective list theories can be found. It is (as we will see) only in this area that the distinction between the desire theory and the objective list theory is a real dichotomy, i.e. it is not just that they purport to answer the same question (or questions); their respective answers are also mutually exclusive.

So the question arises: If this is so, what question (or questions) are the two theories trying to answer? What claims do they make, and what is it that they (essentially) disagree about? To find an answer to this question, let us first take a further look at how the two theories have been characterized in the literature. Here are some examples of how the Objective List Theory has been characterized[2]:

According to Parfit (1984) (who seems to have invented the name of the theory), the central feature of the Objective List Theory is (as we have already seen) the idea that "certain things are good or bad for us, even if we do not want to have the good things or avoid the bad things (p 4)[3]".

In Scanlon's (1993) terminology, Objective List Theories (or, as he prefers to call them, Substantive Good Theories) "are theories according to which an assessment of a person's well-being involves a substantial judgement about what things make life better, a judgement which may conflict with that of the person whose well-being is in question" (p 188); theories which "unlike desire theories, /.../ are based on substantive claims about what goods /.../ make life better" (ibid., p 189). And he also points out that on the Substantive Good Theory, a (correct) evaluation of a person's life is "not wholly dependent on" that person's

---

[1]It is important to note that in the present context (e.g., when I contrast the desire theory with the objective list theory), I will (with almost all other writers) have the *object interpretation* of the desire theory in mind. I will also (for the sake of simplicity) have so-called *actual* desire theories in mind.

[2]And note that since the distinction between the desire theory and the objective list theory is a dichotomy, what follows is also an indirect characterization of the desire theory.

[3]Which of course implies that some kinds of hedonism (viz. quality hedonism) are to be regarded as "objective list theories".

"tastes and interests" (ibid., footnote 10, pp 188-189).

On Kagan's (1992) view, Objective List Theories (or, as he sometimes prefers to call them, Objective Theories) are theories which "hold that various things are objectively good for a person to have, whether or not he realizes it, and whether or not he desires it. Being well-off is simply a matter of one's having the various objective goods. These might include not only pleasure, but also, for example, friendship, fame, knowledge, or wealth. The list of objective goods is, of course, a matter of dispute, but there is no obvious reason to think it would be restricted to kinds of mental states" (p 170).

In Griffin's (1986) terminology, an objective account (of well-being) is an account "that makes well-being independent of desires" (p 32): "An objective-list approach says that a person's well-being can be affected by the presence of certain values (which it lists) even if they are not what he wants" (ibid., p 33). When these "objective" values "appear in a person's life, then whatever his tastes, attitudes, or interests, his life is better" (ibid., p 54).

Here, it is important to notice that an objective list theorist need not claim that the value of a life is wholly independent of desires, i.e. "that we can measure changes in a person's well-being just by the amount that he realizes objective values" (ibid., p 54), or that "being well-off is simply a matter of one's having the various objective goods" (Kagan (1992), p 170). An objective theory need not be "pure" or "simple" (in this sense): it may (as Scanlon (1993) points out) well allow for the possibility that what final value a life has for the person who lives it is *in part* dependent on what his preferences and desires are:

> As I see it, according to a desire theory, when something makes life better this is *always* because that thing satisfies some desire. Substantive good theories can allow for the fact that this is *sometimes* the case - it is sometimes a good thing simply to be getting what you want - but according to those theories being an object of desire is not in general what makes things valuable (p 190, my italics).

What, then, does all this amount to? Exactly how does the Objective List Theory (Substantive Good Theory) differ from the Desire-Fulfilment Theory (i.e. the object interpretation of the theory), and (above all) on what issue (or issues) do the two theories disagree?

Let us first look at the (actual) desire theory. The central claim of this

theory is the idea that what is nonderivatively good (and bad) for a person is "wholly dependent on" what (intrinsic) desires (and aversions) he has. If we take a close look at the quotations above (and add a little extra knowledge), we can see that this central claim can (in the case of goodness) be divided into two claims (claims that should not be regarded in isolation):

(D1)[4] A situation X is nonderivatively good for a person P *if and only if* P has an intrinsic desire that X obtains (and this is a necessary truth), and

(D2) If X is nonderivatively good for P, then this is so *because* P has an intrinsic desire that X obtains. That is, the fact that X is an object of P's intrinsic desire *makes* it nonderivatively good for him.

The objective list theory can (again, in the case of goodness) be characterized as follows:

(O1) It rejects (D1)[5],

(O2) it rejects (D2), and

(O3) it makes (on top of this) substantive claims about what has nonderivative value for a person[6].

Now, to understand what these theories (really) claim, and what questions they (really) purport to answer, it is of utmost importance to find out what (D1) and (D2) (really) mean. So, let us start with (D1): How should the phrase "if and only if" be understood here?

As I see it, the statement "X is nonderivatively good for P if and only if

---

[4]It is important to note that the notation in this appendix differs from the notation which is used throughout the rest of the essay.

[5]That is, it allows for the possibility that there are things that are good for a person P even though they are not wanted by P, and that there are things that are wanted by P that are not good for P. However, the objective list theory also allows for the possibility that there are persons for whom "the good" and "the wanted" happen to coincide.

[6]However, this may not be a complete characterization of the objective list theory, and for the following reason: There are other versions of the desire theory than the actual desire theory (e.g., so-called idealized versions of the theory), versions that have exchanged (D1) and (D2) for other claims. And we do not have a complete characterization of the objective list theory until we know whether it accepts or rejects these claims.

P intrinsically desires that X obtains" can be interpreted in two different ways, viz. (i) as a formal criterion of goodness-for, and (ii) as a metaphysical (or ontological) statement about the nature of value-for.

(i) First, the statement can be regarded as a (general) statement about how it should be determined what it is that has value for a person, i.e. as a formal standard by which it can be judged (or decided) whether or not something is good for a person. To interpret (D1) in this way is to regard it as an answer to the question "How do we determine what has value for a person?" rather than as an answer to the substantive "what is nonderivatively good for a person?".

(ii) But (D1) might also be regarded as a metaphysical (or ontological) statement about the nature of value-for, as a claim about what value-for *is*, rather than as a claim about what it is that *has* value, or about how we should determine what it is that *has* value. If (D1) is regarded in this way, then what it states is that to be nonderivatively good for P *is* to be the object of some of his intrinsic desires, or alternatively, that the nonderivative value of a situation X for a person P is somehow *constituted by* the fact that P intrinsically desires that X holds. On this interpretation, (D1) is viewed as an answer to the metaphysical (or ontological) question "What is it for a thing to have nonderivative value for a person?", or alternatively, "If X is nonderivatively good for P, then what sort of relation is this, e.g., is it a *sui generis* relation or not?".

To understand (D1) in this way is (I think) to understand it as a version of what Bergström (1990) calls ontological naturalism: The (evaluative) "property" of having nonderivative value for a certain person is regarded as identical with the empirical (or natural) "property" of being intrinsically desired by this person, and the evaluative fact that a situation X is nonderivatively good for a person P is regarded as identical with the empirical (natural) fact that P intrinsically desires that X obtains.

Now, it is quite clear that the ontological interpretation (ii) of (D1) implies the evaluative interpretation (i). But Bergström (1990) also suggests (on pp 56-57) something stronger, viz. that (D1:ii) is identical with the conjunction of (D1:i) and (D2)[7]. Personally, I do not agree with

[7]What he really claims in his book is that the following two interpretations of utilitarianism are identical with each other, viz. (i) utilitarianism as a criterion of rightness (what he calls U) *plus* the idea that an action is right *because* it has better consequences than the alternatives (the counterpart to (D2)), and (ii) utilitarianism

Bergström, and for the following reason: As (D2) suggests, there is (on the desire theory) an asymmetry between the fact that X is good for P and the fact that X is desired by P: X is good for P *because* P desires X, but not vice versa. Therefore, we should not regard the two facts as identical[8]. Bergström claims that this is a bad argument, but as far as I can see, he does not give us any reason for accepting this claim[9].

So, *if* (i) and (ii) are different interpretations of (D1), which one is to be preferred? I think we should go for (i); it is more in line with the spirit of the desire theory, and it is somehow more in harmony with (D2). But remember that (ii) implies (i), so even if we do not interpret (D1) as a formal criterion of nonderivative value-for, it implies such a criterion. That is, it is quite clear that it is in virtue of (D1) that the desire theory is (in part) an evaluative theory, viz. a certain kind of formal theory[10].

Let us now move on to the second essential difference between the two theories, viz. the disagreement about the validity of (D2). As in the case of (D1) and (O1), it is not entirely clear how (D2) and (O2) should be understood. The source of the unclarity is this: When it is claimed

---

as an ontological thesis (what he calls T*), i.e. the idea that the property rightness of an action is identical with the empirical, natural property to have better consequences than the alternatives.

[8]Cf. Hare's (1952) criticism of metaethical naturalism.

[9]Cf. what Sumner (1996) writes (on p 16): "A theory of the nature of welfare must /.../ be formal. It must tell us what it is for someone's life to go well or badly, or for someone to be benefited or harmed. In order to do so it must provide the appropriate relation to complete such formulas as '*x* benefits *y* if and only if *x* stands in relation *R* to *y*'". This suggests that Sumner is either unaware of the difference between (i) and (ii), or he takes it for granted that an answer to "How do we determine what has value for a person?" can not be plausible unless it is also an answer to the metaphysical (or ontological) "What is the nature of prudential value?", i.e. that the former question must (so to speak) be approached "through" the latter question.

[10]From this, we can conclude that it is (O1) and (O3) that characterize the objective list theory *qua* evaluative theory. This is not very informative, however; all it tells us is that the theory is not a formal theory of a certain sort, and that it is (in part) a substantive evaluative theory. This does not imply that objective accounts can not be formal, however, i.e. to call the objective list theory "the substantive good theory" (like Scanlon does) can be rather misleading. In any case, the fact that the desire theory is merely formal explains why the two theories may well agree on what is good for a certain particular person (cf. also note 5#); what they could never agree on is *why* a certain situation good for a certain person, i.e. on what it is that *makes* it good for this person.

that "if something is good for a person, this is always *because* this thing satisfies some desire", or that "being an object of a person's desire is what *makes* something valuable for this person", it is not clear how key terms like "because", and "makes", should be interpreted here. So, we need to know how these key terms can (and should) be interpreted.

As I see it, (D2) can (depending on how these key terms are interpreted) be interpreted in two different ways, viz. (a) as an ontological (or metaphysical) statement about "the source" of prudential value, or (b) as a statement about how value-for-claims can (and should) be justified.

(a) On the first interpretation, (D2) is regarded as a metaphysical (ontological) statement about "the source" of value-for-P. If (D2) is viewed in this way, then it is identical with a certain form of *subjectivism*, viz. the idea that desire is (ontologically) prior to value-for, or alternatively, that value-for-P is ontologically dependent on P's desires (e.g., if P would not have any intrinsic desires, then there would be no such thing as nonderivative value-for-P)[11].

This would make (O2) identical with a certain kind of "objectivism", viz. with the idea that there are objective prudential values, i.e. facts about what is good and bad for a person P which hold independently of what P or anyone else thinks or feels about the matter[12]. (This also suggests that certain things could be good or bad for us, even if we had no desires at all)[13].

---

[11]It is important to notice that this view is compatible with realism, i.e. that it does *not* claim that value-for-people-statements lack truth-value. What it claims is rather that *if* such statements have truth-values at all, then these truth-values are mind-dependent in a certain way, viz. they are dependent on the truth-values of statements of the form "X is intrinsically desired by P". Notice that this type of subjectivism is similar to, but not identical with the ontological naturalism above. It is also similar to, but (again) not identical with, the "projectivist theory of value", a theory which claims that the so-called values of things are projections of attitudes we take towards them, i.e. that values (if they can be said to exist at all) are "products of our minds". (But notice that projectivism is normally regarded as a theory about aesthetic value (cf. the well-known "beauty lies in the eye of the beholder") or value-period rather than as a theory about value-for).

[12]Here, it is important to see that this form of objectivism does not imply that what has value for P is independent of *what P is like*, e.g., what is good for P may (on the objectivist view) well be dependent on what his needs or abilities are.

[13]It might be argued that objectivism in this sense (which is compatible with naturalism, i.e. which differs from the view that value-for is a *sui generis* relational property) is identical with realism. But what is "realism" supposed to mean here? Well, it can not refer to the idea that value-for-people-statements have truth-

(b) The second way in which (D2) and (O2) can be understood are as statements about how (substantive) value-for-statements can (and should) be justified. To interpret (D2) and (O2) in this way is to regard them as answers to the question "How do we justify claims (or beliefs) about what has nonderivative value for a person?". Viewed as answers to this question, (D2) claims that the validity of the statement "P has an intrinsic desire that X" is a good reason for accepting the statement "X is nonderivatively good for P" as valid, while (O2) denies this. Here, it is important to see that (O2) is not a rejection of the (plausible) idea that statements of the form "X is good for a person P" must be justified by referring both to what X (the object) is like and to what P (the subject) is like. All that is rejected by (O2) is the idea that the relevant features of the subject (in this context) is what he intrinsically desires: it allows for the possibility that needs, talents, or abilities are more relevant.

The question of justification (of value-for-claims) to which (D2) may constitute an answer can be specified as follows: What kind (or kinds) of subject-oriented reasons[14] can be given for substantive evaluative claims of the form "it is nonderivatively good (or bad) for a certain *particular person* that facts of type X obtain"? What constitutes (in this context) a good argument? This is the question which a subjectivist tends to regard as the *real* problem of justification. An objectivist, on the other hand, tends to regard the question "What kind (or kinds) of subject-oriented reasons can be given for substantive evaluative claims of the form "it is nonderivatively good (or bad) for *all human beings* that facts of type X obtain"?" as the *real* problem of justification. The reason why the subjectivist and the objectivist tend to differ in this way is that objectivists are more inclined to accept the "universalist" (or "generalist") idea that whenever a fact of a certain type has nonderivative value for a particular person, then this is so because this person is a certain type of creature (e.g., a human being), and because all facts of this type have nonderivative value for all creatures of the type in question. It is important to point out that an objectivist *need not*

---

values (mind-independent or mind-dependent); this type of realism is compatible with subjectivism too. So, it seems that objectivism above can only be identical with realism if we define the two positions in the same way, i.e. as the idea that value-for is a relational property which is independent of our attitudes and concerns.

[14]Concerning how the term "subject-oriented" should be understood in this context, cf. e.g., appendix B, in the section on supervenience.

be a "universalist", however, i.e. that objectivism is compatible with "relativism"[15].

But why is this? After all, it seems that the objective list theory implies, in virtue of its objectivist component, a certain kind of "universalism", viz. the idea that what is good for me is also good for you, and vice versa, or more specifically, the idea that if a certain situation is nonderivatively good for a certain person, then there is a general description of this situation such that the claim "Situations of this type are nonderivatively good for everyone" is true (Cf. Scanlon (1993), according to whom "[t]he term /.../ 'objective' suggests a kind of rigidity (as if the same things must be valuable for everyone)" (p 188)). So why do I claim that objectivism does not really imply "universalism"? Well, this is my reason: The only reason why most objective list theories are also "universalist" theories are that they assume that statements of the form "all situations of type X are good for P" must be (in part) justified by referring to some feature of P that he shares with all human beings, e.g., to some universal human need, or a common human nature. An objectivist need not attempt to justify goodness-for-P-statements in this "generalist" way, however; instead, he may refer to features of P which may well be unique to P (but remember that he must not refer to P's desires), e.g., to P's individual potential (talents or abilities), or to his individual needs. In short, objectivism is compatible with "relativism"[16].

---

[15]Where the term "relativism" refers to the idea that what is good for me may not be good for you, and vice versa, or more precisely, to the idea that there are types of situations which have nonderivative value for some people but not for others, even if these situations are described on as high a level of generality as possible (or required). However, the term "relativism" normally refers to a certain ontological view (a view about truth), or to a certain epistemological view (a view about justification), and I am (for this reason) somewhat reluctant to use it in this way. In fact, the present use of the term isn't even mentioned by Bergström (1990), who lists seven different uses of the term (on pp 112-117).

[16]It might be argued that all relativist theories of goodness-for (full-fledged or moderate) are *only apparently* relativist, however, that if we look at how they justify goodness-for-P-statements, we see that they are ultimately of a universalist nature. For example, suppose there is this relativist who claims that it is good for P to engage in intellectual activity, but not for Q. Now, if we ask why this is so, we will be told that it is because P is different from Q in some relevant respect, e.g., because P has some talent that Q does not have. And (so the argument goes), for this type of justification to be valid, we have to assume that it is good for everyone to actualize his or her potential. Now, the main problem with this argument is that the general idea that it is good for us to actualize our respective potentials can (as I see it) only be plausible if it is viewed as a formal claim rather than as a substantive claim. Or alternatively put, it is not plausible to attribute

Before we end the section on what question (or questions) that the desire theory and the objective list theory purport to answer, let us first say a few words on the issue of *"fallibilism"*. According to some philosophers, the objective list theory claims that our value-for-judgements may be mistaken, e.g., Scanlon (1993), who tells us that Substantive Good Theories are "theories according to which an assessment of a person's well-being involves a substantive judgement about what things make life better, *a judgement which may conflict with that of the person whose well-being is in question*" (p 188, my italics), or Kagan (1992), who tells us that Objective Theories "hold that various things are objectively good for a person to have, *whether or not he realizes it*" (p 170, my italics). Here, it is important to see that they do not just have particular evaluations in mind, but also more general evaluations. The claim is not just that we all may (on the objective list theory) be mistaken about what nonderivative value that a particular situation has for us (e.g., whether it is good or bad), or about how well off we are (on the whole), but also that we may be mistaken about what types of situations that are good and bad for us, or about what it is (in general) that makes our lives go better or worse.

But is this something which a desire theorist must deny? As far as I can see, a Desire Theorist need not claim that we are infallible as "particular evaluators", especially not when the object of the particular evaluation is as complex as a whole existence-at-a-certain-time[17]. So what

---

nonderivative value to the circumstance that one's potential is actualized, but it may well be plausible to claim that it is nonderivatively good for P to engage in intellectual activity because this constitutes the realization of some talent that he has.

Before we leave the issue of relativism vs. universalism, it is worth noting that we have (in the above) only been concerned with one type of relativism, viz. "person-relativism", i.e. the idea that substantive prudential values may be relative to individual persons ("what is good for me may not be good for you"). It is, however, important to point out that there are other possible types of relativism, e.g., prudential values may also be relative to cultures (personal development may be good for "modern people" but not for "traditional" or "primitive" people), and they may be "time-relative": If a certain prudential value is relative to a certain person, it may either be relative to this person "over time" (if something is good for me now, it will also be good for me later, and vice versa), or to this person at a certain time (what is good for "me-now" may not be good for "me-then", or vice versa).

[17]What a Desire-Fulfilment Theorist must claim is (a little oversimplified) that the value of a person's life is a function of to what extent his relevant desires are fulfilled. But this function may (at least if we have the non-global versions of the

about the general case? Must a desire theorist really reject the idea that we may be mistaken about what types of situations that have non-derivative value for us? This is a tricky question. There are (of course) two obvious senses in which a person can, on the desire theory, be mistaken about what is good for him; first, he may not know what his intrinsic desires are and/or how strong these desires are (in practice, it may be just as difficult to find out what one really wants as it is to discover the correct list of "objective values"!), and second, he may not realize that the desire theory is valid (e.g., he may, from his (immanent) perspective, believe that we desire something because it is good for him, rather than the other way around). But if we disregard these ideas, it seems that the (actual) desire theorist is not much of a "fallibilist" about goodness-for. After all, he accepts the idea of "the sovereign subject" (cf. section 5.1.1).

To conclude appendix D: We have seen that the three central questions of this essay are not the only important questions of prudential value (and well-being). For example, there are also questions like "What is the nature of nonderivative goodness-for? What is it for a situation to have nonderivative value for a person; what does the relation goodness-for consist in?"; "What is the source of nonderivative goodness-for? If goodness-for-statements can be true, what is it that makes them true?"; "To what extent can we be mistaken about what it is (in general) that makes our lives go better or worse?"; and "Are there any universal prudential values, i.e. are there any valid substantive claims of the form 'All facts of type X are nonderivatively good (or bad) for all human beings (at all times)'?".

But most importantly, there are a number of (substantive) questions of justification: questions of how (exactly) value-for-claims can (or should) be justified. Depending on what type of value-for-claim we have in mind, there are two possible problems of justification, viz. how should substantive claims of the forms (1) "All facts of a certain type X are nonderivatively good (or bad) for a particular person P", and (2) "All facts of a certain type X are nonderivatively good (or bad) for all human beings (at all times)" be justified?[18]

---

desire theory in mind) be too complex for some of us to grasp.

[18]The reason why I ignore the question of how particular claims of the form "a particular fact X is nonderivatively good (or bad) for a particular person P" should

Which of the two problems that is most central depends on whether there are universal prudential values, i.e. on whether we can assume that there are valid substantive claims of the form "All facts of a certain type X are nonderivatively good (or bad) for all human beings" (and that such claims can be justified).

If it can be assumed that the important prudential values are universal, the central question is (2): How can (should) claims of the form "all facts of a certain type X are nonderivatively good (or bad) for all human beings (at all times)" be justified (or refuted)? What kind (or kinds) of reasons can be given for (and against) such claims? What constitutes (in this context) a good argument? Or alternatively, how can a general (substantive) conception of prudential value be justified (or refuted)?

If it can *not* be assumed that the most important prudential values are universal, the central question is (instead) (1): How (exactly) can semi-general substantive claims of the form "All facts of a certain type X are nonderivatively good (or bad) for a particular person P" be justified (or refuted)?

Now, I have already assumed (e.g., in chapter 1) that there are universal prudential values, i.e. in this essay, it is (2) which will be regarded as the central problem of justification. This question will not be treated separately, however, but only in connection with (I)-(III).

---

be justified is that I have made the "generalist" assumption that whenever a particular fact X has nonderivative value for a particular person, then there exists some description of X such that all other particular facts that fall under the same description (that are of the same type) have nonderivative value for P too.

Appendix E
Hedonism and time
On the issue of duration

If we take the duration of our experiences into account, we can attribute the following view to the hedonist (in answer to (II); cf. note 52 on p 31):

(H5') The (positive or negative) value of an experience for the person who has it is a function of two things only, viz. (i) how pleasant or unpleasant it is, and (ii) how long it lasts (or how long it appears to last). That is, the degree to which a pleasant experience is nonderivatively good for the experiencing subject is a function of its degree of pleasantness and its duration, and the degree to which an unpleasant experience is nonderivatively bad for the subject is a function of its degree of unpleasantness and its duration. The function in question can be characterized as follows:

(i) The more pleasant an experience is, the (nonderivatively) better it is for an experiencing subject to have it, i.e. the higher is its prudential value (and so on; cf. (H5), e.g., in section 1.2). For example, it is always better for a person to have a more pleasant experience of a certain duration than to have a less pleasant experience of the same duration, and it is always worse for a person to have a more unpleasant suffering of a certain duration than to have a less unpleasant suffering of the same duration.

(ii) If two pleasant experiences are equally pleasant, but of different duration, it is (nonderivatively) better for the subject to have the experience that lasts longer, and if two sufferings are equally unpleasant, but of different duration, it is better for the subject to have the shorter experience. If we want to incorporate the idea that it is "subjective duration" that matters, we only have to replace "longer" ("shorter") with "subjectively longer (shorter)" or "appears longer (shorter) to the experiencing subject".

As it stands, this last claim is not very precise. To make it more precise, the pure hedonist would most probably appeal to the following *principle of multiplication*:

If P_n is a measure of the degree of pleasantness (unpleasantness) of an experience E_n, and if D_n is a measure of its duration, then how good (or bad) it is for a certain person to have a certain experience E_1 depends on one thing only, viz. how large the product (P_1 x D_1) is. For example, if E_1 and E_2 are both pleasant, it is nonderivatively better for a person to have E_1 than to have E_2 if and only if the product (P_1 x D_1) is bigger than the product (P_2 x D_2)[1].

However, in order to make (H5') sufficiently clear and precise, we also need to know how the term "duration" should be understood, viz. whether it is to be understood in the objective sense or in the subjective sense. Which of the two possible senses of the term makes (H5') most plausible? Is it objective time (clock time) or subjective time (felt time) that matters? Before we turn to this question, let us first look at the different ways in which the term "duration" can be understood here.

Objective duration

The objective duration of an event (e.g., an experience) is the duration that is measured with clocks. That is, an event lasts longer in the objective sense than another event if and only if it lasts longer according to clock time. And if a certain event lasts for two minutes and another event lasts for four minutes, then the objective duration of the second event is twice as long as the objective duration of the first event.

Two conceptions of subjective duration: the atomistic-summative and the holistic

The subjective duration of an experience (or experienced event) has to do with how long it appears to last (or how long its parts appear to last) to the experiencing subject. But what exactly is this supposed to mean? Or alternatively, how exactly should we determine which of two experienced events that lasts longer in the subjective sense? Well, there are at least two major ways in which this question can be answered,

---

[1]If this kind of multiplication is to make sense in all kinds of cases, it is necessary that the duration of an experience and its degree of pleasantness (unpleasantness) are both measurable on ratio scales (and this is a dubious assumption indeed). Such a high degree of measurability is not required in all cases, however, e.g., it is sometimes sufficient that either duration *or* pleasantness (unpleasantness) is measurable on a ratio scale, while the other variable is merely ordinally measurable.

viz. in an "atomistic way" or in a more "holistic way". Edgeworth[2] and Tännsjö (forthcoming) are examples of philosophers who adopt the atomistic conception of subjective time, and my own proposal is an example of a more "holistic" conception of subjective time.

My own proposal is based on the idea that all appearances are perspectival. In the case of space, this means that if a physical object appears in a certain way to a certain subject, it always appears to him in this way as viewed from some spatial perspective or other. That is, when we say that a certain object appears in such-and-such a way to a certain subject, what we have in mind is not the subject *simpliciter*, but the subject *qua* located at a certain spatial position. Now, in the case of experienced duration of experienced events, the idea is (rather) that if it appears to a person P that an experienced event E lasts so-and-so long, then it always appears this way to P from some *temporal perspective* or other. So, when we say that a certain experienced event appears to last so-and-so long to a certain subject, what we have in mind is the subject *qua* temporally located, i.e. the subject-at-a-certain-(objective)-time (or P-at-t). But the appearance of duration might vary with the temporal perspective of the subject, e.g., a certain event might appear to last relatively long when it is going on, but relatively short shortly afterwards[3].

This is the idea on which my "holistic" proposal is based: How long the subjective duration of a certain experienced event is for a person P depends on how long it (the event as a whole) appears to last (to P).

This formulation will not do, however, and for the following reason: First, how long a certain event appears to last (to P) might vary with P's temporal location (e.g., the event might appear to last longer to P-at-t_1 than to P-at-t_2), and second, there is no one temporal perspective that can be regarded as "privileged".

If the basic "holistic" intuition is combined with these "perspectival" insights, we might arrive at the following "holistic" and "perspectival" conception of subjective duration: Two experienced events have the

---

[2]Edgeworth's view is presented in Tännsjö (forthcoming), on pp 67-69.

[3]It is worth pointing out that we tend to "measure" subjective time in minutes, hours, days, and so on, e.g., as in "it appears to me that this show has been going on for more than three hours". It is also worth noting that judgements of this type seem to be influenced by many things, e.g., like feelings of hunger, perceptions of light and darkness, and how bored or stimulated we are.

same subjective duration for a person P if and only if there is no point in time at which the two events appear (to P) to be of different duration[4]. And an experienced event E1 lasts longer (in the subjective sense) than another experienced event E2 if and only if there is at least some point in time at which E1 appears to last longer, *and* there is no point in time at which E2 appears to last longer[5].

As an illustration of this idea, consider the following two examples:

(i) Suppose that a certain person is totally absorbed in writing a book, or in listening to music. Now, to be absorbed is to be in a kind of timeless state, and the person would not (at any time) experience any difference between being absorbed for three hours and being absorbed for yet another hour. If this is so, we can say that the two experiences have the same subjective duration for this person.

(ii) Suppose that a person goes to a two hour long concert, to listen to a symphony that he knows by heart. Now, consider the following two cases: In the first case, the person enjoys the concert for two hours and believes that he has done so, and in the second case, he enjoys the concert for half an hour, has a black-out that lasts for an hour, regains consciousness, enjoys the concert for another half hour, and then believes that he has enjoyed it for two hours. On the holistic conception of subjective time, these two experiences have the same subjective duration.

The atomistic and summative conception of subjective duration is "based on" the idea that subjective time can be divided into small units; units that can not be further divided. There are different views on what this smallest unit is, e.g., it might be the least noticeable difference in time (the least difference in time that can be directly discriminated), as Edgeworth suggests[6], or it might the least sub-noticeable difference in time (the least difference that can be indirectly discriminated), as Tännsjö (forthcoming) suggests. If we call this unit a "subjective

---

[4]No point in time within a certain interval, that is: the idea is that we can ignore those times which occur before the last of the events, as well as those times which occur long after the last event. This is about as precise as I can get, and it is worth noting that this is but one respect in which my conception of subjective duration is very rudimentary and incomplete.

[5]This means that if E1 appears to last longer at certain times, while E2 appears to last longer at other times, then we have an insurmountable problem of measurability: How do we, in cases like this, rank experienced events with respect to subjective duration?

[6]See note 2.

second", the atomistic-summative conception of subjective duration can be characterized as follows: The subjective duration of an experienced event is a matter of how many subjective seconds it lasts (i.e. subjective time is, on this view, measurable on a ratio scale)[7]. For example, that an experienced event E1 lasts longer in the subjective sense than another experienced event E2 simply means that the number of subjective seconds that passes when E1 goes on is larger than the number of subjective seconds that passes when E2 goes on. And if E1 lasts for 100 subjective seconds while E2 lasts for 50 subjective seconds, then the subjective duration of E1 is twice as long as the subjective duration of E2.

Three different versions of pure hedonism

We have now seen that there are at least three possible senses of "duration", viz. objective duration, subjective duration of the holistic kind, and subjective duration of the summative kind. Depending on which of these senses of "duration" that the hedonist has in mind, there are different versions of the hedonistic theory, e.g., of (H5'):

If the pure hedonist has objective duration in mind, we get:

> **(H5'O)** The value of an experience for the person who has it is a function of two things only, viz. (i) how pleasant or unpleasant it is, and (ii) how long its objective duration is. And since objective duration is measurable on a ratio scale, we can appeal to the multiplication principle, viz. in the following way: If P_n is a measure of the degree of pleasantness of an experience E_n, and if D_n is a measure of its objective duration, then the value that it has for a certain person to have a certain experience E_1 is proportional to how large the product (P_1 x D_1) is. For example, if two pleasant experiences are equally pleasant, and if one of them lasts twice as long as the other, we can conclude that it is twice as good for the experiencing subject to have the longer experience.

If the hedonist has (instead) the atomistic (and "summative") kind of subjective duration in mind, we get:

---

[7]Notice that atomist assumes from the very beginning that subjective time is measurable on a ratio scale, i.e. the problem of whether subjective time is ordinally measurable is (I think, deliberately) bypassed.

(H5'SA) The value of an experience for the person who has it is a function of two things only, viz. (i) how pleasant or unpleasant it is, and (ii) how long its "atomistic-summative subjective duration" is. And since this type of subjective duration is (just as objective duration) measurable on a ratio scale, we can appeal to the multiplication principle in exactly the same way as we did under (H5'O) above.

If the hedonist has the holistic kind of subjective duration in mind, if he (instead) incorporates this version of the idea that "subjective duration is what matters" into his theory, we get:

(H5'SH) To the extent that it is possible to determine how valuable it is for a certain person to have a certain experience: The value of an experience for the person who has it is a function of two things only, viz. (i) how pleasant or unpleasant it is, and (ii) how long it appears to last, or more precisely, how long its "holistic subjective duration" (in the sense this was given above) is. For example, if two pleasant experiences, E1 and E2, are equally pleasant, then it is nonderivatively better for a person to have E1 than to have E2 if and only if E1 is subjectively longer (in the holistic sense) than E2, i.e. (roughly) if and only if it appears (as a whole) longer than E2 to the experiencing subject.

So, the question arises: Which of these three versions of hedonism is the most plausible evaluative theory; (H5'O), (H5'SA), or (H5'SH)? Or alternatively put, which of the three possible senses of the term "duration" should the hedonist accept?

Why it is the holistic kind of subjective duration that matters

Personally, I tend to believe that the most ethically relevant kind of duration of experiences is subjective duration of the holistic type: what matters to an experiencing subject is how long his experiences appear to last to him. That is, as far as the issue of duration is concerned, the most plausible version of pure hedonism is (H5'SH), e.g., if we have objective duration or "atomistic and summative subjective duration" in mind, then it is *not* always the case that the longer a pleasant experience lasts, the better it is for the experiencing subject, and neither is it always the case that the longer a suffering lasts, the worse it is for the experiencing subject. These are my "arguments" for this view:

(i) The first argument purports to eliminate (H5'SA) from the competition: The reason why we are interested in subjective duration at all is because we suspect that how long a pleasant (or unpleasant) experience actually lasts is (in this context) of less significance than how long it appears to last. For this reason, a plausible conception of subjective duration of experienced events cannot totally ignore how long an experienced event (as a whole) appears to last to the experiencing subject, and neither can it ignore that appearances of duration are perspectival, i.e. that they might vary with the temporal perspective of the subject. The atomistic-summative conception ignores both these things, however, and it is (therefore) not a plausible conception of subjective time. In fact, this conception is not really a conception of subjective time at all. It is true that the unit is constructed on subjective grounds, but once the unit has been constructed, subjective time is regarded in an entirely objective manner, i.e. the subject's own perspective is transcended entirely. In short, what the atomist has constructed is simply a new kind of clock; a clock that measures subjective time from an objective (impersonal) and atemporal standpoint. And since there is no reason whatsoever for believing that the subjective duration measured in this way has anything to do with how things appear to us on the macroscopic level (the connection between the two kinds of subjective duration might even be weaker than the connection between objective duration and felt (macroscopic) duration!), why doesn't the atomist just stick to proper objective time? (Especially since there is no reason whatsoever to believe that subjective time is an additive quality). In short, it seems that the atomistic-summative conception of subjective duration is not really an alternative to objective duration, and even if it is, I can't see how one would decide which of the two "durations" that is more ethically relevant.

(ii) So, how are we to decide between (H5'SH) and (H5'O)? Well, we have to resort to the kind of particular cases (real or imagined) that the two theories would disagree on, e.g., cases like the following ones:

(a) "The holistic subjectivist" (H5'SH) claims that if two experiences are equally pleasant, and if there is no point in time at which a person would experience any difference in duration between two experiences, then they are equally good (or bad) for this person. This suggests that it does not really matter to a person whether he is totally absorbed in a certain activity for four hours or for eight hours (assuming that absorp-

tion is a kind of "timeless state", and that he would not notice any difference). "The objectivist" (H5'O) would deny that this is so. Instead, he would claim that it is better for a person to be absorbed for eight hours than to be absorbed for four hours. He would also claim, on top of this, that it is *twice as good* for the person to be absorbed for eight hours than to be absorbed for four hours (this is nothing but an application of "the rigid additive thesis with respect to duration").
(b) "Holistic subjectivism" (H5'SH) implies (it seems) that it is not non-derivatively worse for a person to be tortured for two hours and to feel as if he has been tortured for two hours than it is for him to be tortured for half an hour, faint, regain consciousness, be tortured for another half hour, and then feel as if he has been tortured for two hours. "Objectivism" (H5'O), on the other hand, does not just imply that the objectively longer experience is worse, but also that it is *twice as bad* as the shorter experience.

On my view, (H5'SH) seems more plausible than (H5'O). "Objectivism" is not a plausible view: The claim that it does not matter at all how long our pleasures and sufferings appear to last is not plausible, and neither is the idea that it is twice as good for a person to be absorbed for eight hours than to be absorbed for four hours, even if the person himself would not notice any difference between the two experiences[8]. This does not mean that (H5'SH) is a satisfactory view, however. As I see it, the whole temporal issue is rather messy (which might suggest that it is a fruitful area for future research).

Does duration (per se) matter at all?

So, the basic intuition here is that it matters how long our pleasures and sufferings appear to last. But does this idea really imply (H5'SH), i.e. that the value of an experience is (in part) a function of how long it appears to last? Well, not necessarily. There is (it seems) another way in which appearances of duration might matter, viz. by being (in a certain way) "embedded" in our experiences (and if this is the only way in

---

[8]The objective view seems more plausible if we have value-period (the value of the world) in mind, however, e.g., it is neither unintelligible nor implausible to say that it is twice as good-period that a certain person is happy for two days than it is that he is happy for one day. The reason for this is simple: If it is value-period we are concerned with, we should transcend all limited temporal and personal perspectives, and instead adopt a more atemporal and impersonal perspective.

which temporal appearances matter, duration might not matter at all).

Consider the following line of reasoning: What duration (objective or subjective) a certain experience has is *not directly relevant* to the question of what value this experience has for the experiencing subject. It is true that in most cases, a longer suffering is worse than a shorter suffering. But the explanation for this is that the subjective duration of an experience is almost always "built into" this experience. For example, my present suffering or enjoyment is (in part) based on thoughts about the past and the future. The reason why a two hour long visit to the dentist makes me suffer more than a visit that is only one hour long is probably this: If I suffer longer than a certain limited time, I will most probably start to think things like "now this has been going on for so-and-so many minutes; when will it ever end?" and "when this is over, everything will be just great", and this will (in turn) tend to increase my suffering at that very moment. And if I think "now it will soon be over", I may suffer less. But if I don't get absorbed at all in ideas about the past or about the future, what difference will it make how long I suffer?[9] Or alternatively put, if someone would live entirely in the present, if he would have no consciousness of time, would it really be intelligible to say that a longer suffering is nonderivatively worse *for him* than a shorter suffering? That is, duration per se might not (in the context of value-for, that is) matter at all; objective duration might not matter unless a sense of it is incorporated into present experience, and subjective duration might not matter unless it is incorporated into present experience.

There is another argument that can be given for the idea that as long as we have value-for in mind, duration does not really matter, viz. the following one: The idea that duration matters can be expressed as follows: "If X and Y are two pleasures of different duration (objective or subjective), it is nonderivatively better for a person P to have the pleasure that lasts (objectively or subjectively) longer". But is this really an intelligible idea?

It has already been suggested (e.g., in appendix C) that all value-for-P is value-for-P-at-*t*, i.e. that it does not make sense to talk about value for P-over-time. Or alternatively put, that if a particular value-for-

[9]And similarly for pleasure. For example, if I have pleasant dreams every night, is it really nonderivatively better for me the longer these dreams last? (Is it good for me at all to have pleasant dreams if I forget all about them afterwards?).

statement of the form "X is (was, or will be) good for P" is fully spelled out, it will take the form "X-between-t_1-and-t_2 is (was, or will be) good for P-at-t_3", where t_1, t_2, and t_3 may (but need not) be identical. It has also been suggested that whether a certain value-for-statement makes sense or not, and whether it is plausible or not, this depends (to a certain extent) on the nature of the temporal relation between X and P, or more precisely, between the times referred to (i.e. t_1, t_2, and t_3). For example, it is obviously both intelligible and plausible to say that it is nonderivatively good for P-at-t that he has a pleasant experience at t, but it makes little or no sense to say that it is nonderivatively good for P-at-t that he had a pleasant experience some time in the past.

Let us now, in the light of this, return to the claim that it is nonderivatively better for a person to have a pleasure of longer duration than it is to have a pleasure of shorter duration. Is this a plausible idea? Does it make any sense at all? The first thing we have to do here is to complete the claim, i.e. by spelling out its (implicit) temporal content. This is how this might be done: "It is nonderivatively better for P-at-t_0 that he has a pleasant experience that goes on between t_1 and t_3 than that he has a pleasant experience that goes on between t_1 and t_2 (where t_1 occurs before t_2, which (in turn) occurs before t_3)". So, how is t_0 related to t_1, t_2, and t_3? As I see it, there are only two interesting possibilities here, viz. either (i) t_0 occurs before t_1, or (ii) it occurs between t_1 and t_2. Let us now see whether any of these two interpretations of the "duration claim" makes any sense, and whether it is plausible.

(i) If the claim is interpreted in the first way, we get the idea that it is nonderivatively better for a person to have a longer pleasure in front of him than it is to have a shorter pleasure in front of him. Is this an intelligible and/or plausible idea? Well, it seems sensible to say that I am (somehow) "in a better position" if I have a four hour long pleasant experience in front of me than if I only have a two hour long pleasant experience in front of me. But in what way is it supposed to better for me-now to have the four hour pleasure in front of me? It is hardly nonderivatively better for me-now. So, is it (then) derivatively better for me-now? Maybe it is not better for me-now at all?[10]

---

[10]So, how should we conceive of derivative value and time? Well, here are two possibilities: (i) We can reject the idea that all derivative-goodness-for-P is derivative-goodness-for-P-at-t, or (ii) we can keep the idea, and claim that X-at-$t1$ is derivatively-good-for-P-at-$t1$ if and only if it makes Y-at-$t2$ possible (etc.), where

(ii) If the claim is (instead) interpreted in the second way, we get the idea that it is nonderivatively better for a person to be in the midst of a longer pleasure than it is to be in the midst of a shorter pleasure. Is this a more intelligible and/or plausible idea? I think not. It is of no interest whatsoever (to me-now) whether the pleasure I experience right now lasts for four hours or whether it lasts for two hours. What matters is how long that is left of it, it is only the future that is of interest to me-now. Or in other terms, it is better for a person to have one hour left of a two hour pleasure than to have half an hour left of a four hour pleasure. In this context, it is (somehow) nonsensical to talk about the value-for of an experience as a whole, especially if we have nonderivative value-for in mind. If a person is in the midst of an experience, then his experience-as-a-whole is partly "in the past", partly present, and partly "in the future", and this is why it doesn't make sense to talk about the value of such a strange entity *for the present person.*

In short, it seems that it does not really make sense to say that it is nonderivatively better for a person to have a pleasure of longer duration than it is to have a pleasure of shorter duration. It might make sense to say that P is in a better position if he has a longer pleasure before him than if he has a shorter pleasure before him, but the betterness referred to here is (most definitely) not nonderivative betterness-for. And it is both sensible and plausible to say that it is, at every objective point in time t, nonderivatively better for P-at-t to feel pleasure at t than not to feel pleasure at t, but from this we cannot conclude that it is better for P the longer his pleasure lasts.

If this is correct, it seems that we can ignore the issue of duration altogether. That is, we can think of hedonism as a theory about what has nonderivative value for people-at-certain-points-in-time. If we do this, we can safely ignore (H5') (the idea that the value of an experience for the person who has it is a function of how pleasant or unpleasant it is, and how long it lasts). Instead, we can remained satisfied with the original (H5), which can now be understood as follows:

> **(H5)** The value that it has for a person-at-a-time to have a certain experience at that time (or rather: to be in a certain concrete conscious mental state at that time) is a function of one thing only,

---

Y-at-t_2 is nonderivatively-good-for-P-at-t_2, *and* P-at-t_1 and P-at-t_2 are the same person. On my view, (ii) is a much more attractive idea.

viz. how pleasant or unpleasant this state is. The more pleasant P's state-at-*t* is, the nonderivatively better it is for P-at-*t* to be in that state.

Appendix F
What conception of desire is most plausible?

The rudimentary conception of desire which was formulated is (as the name suggests) far from complete. So how should it be completed? What must a complete conception of desire be like if it is to satisfy conditions (i)-(iv) on pp 169-172?

Conceptions of desire are normally divided into two main types, viz. phenomenological conceptions and functional conceptions (and as far as I can see, this is an exhaustive classification). So, let us first ask whether the conception of desire we are looking for can be a phenomenological conception.

Phenomenological conceptions of desire

To accept a phenomenological conception of desire is to accept the idea that desires are (essentially) "phenomenologically salient" states, or as Smith (1994) puts it, that "desires have phenomenological content essentially". On this view, a desire is (essentially) something of which we are directly aware, something that is (or can be) felt, something that has a felt intensity. According to Smith, there are two versions of this phenomenological view, viz. a strong one and a weak one. *The strong phenomenological conception* of desire is "the view that desires are, like sensations, simply and essentially states that have a certain phenomenological content" (p 105), and *the weaker phenomenological conception* is the idea that "desires are like sensations in that they have phenomenological content essentially, but differ from sensations in that they have propositional content as well" (ibid., p 108). So, should the desire theorist accept any of these two views? No, he should not, and for the following reasons:

The reason why we should regard the strong phenomenological conception of desires as irrelevant is simple: This view is either hard or impossible to combine with the idea that desires have propositional content (as Smith points out, it cannot provide a plausible epistemology of propositional content of desire). In fact, the strong phenomenological conception is most plausibly regarded as a conception of *desire for* rather than as a conception of *desire that*.

The reason why we should regard the weak phenomenological con-

ception as irrelevant, or alternatively, why we should regard any phenomenological conception of desire (weak or strong) as irrelevant, is twofold: First, the phenomenological notion of desire is too narrow, i.e. it does not satisfy (iii) on pp 170-171. On the use of "desire" that has "rational and moral significance", we are not always directly aware of our desires, and a desire theorist does not have any reason whatsoever to restrict his attention to those desires (in the broad sense, that is) that have felt intensities, or to accept the idea that it is not good for a person to have a desire fulfilled unless this desire is "phenomenologically salient". Second, the phenomenological conception of desire is inconsistent with the "explanatory" or "theoretical" use of terms like "desire", "aversion", and "preference". We are reluctant to eliminate this explanatory use of the terms, however, and this explains why we tend to regard the phenomenological conception as false. For example, the phenomenological view strongly suggests that we are never fallible about the desires we have, and that there are actions that are not manifestations of desires, and these claims are (it seems) simply false.

In short, we should reject all phenomenological conceptions of desire.

Functional conceptions of desire

According to Smith (1994), the alternative to the phenomenological conception of desire is *the dispositional conception of desire*, the view that desires are dispositional states. Or more specifically, to desire that something is or will be the case is (on this view) to have a certain set of dispositions, primarily dispositions to do certain things in certain conditions, but also dispositions to feel certain things in certain conditions. So, is this really the only alternative to the phenomenological conception? I think not. On my view, Smith's dispositional conception of desire is only one of several possible versions of a broader and more general view, a view which we can call *the functional conception of desire*. This is (roughly) the idea that desires are essentially states that have a certain functional role, or more specifically, that desires are (essentially) motivational states, states that are related to action in a certain way (which does not prevent them from being essentially related to other things as well, however). An example of another version of this functional conception is Brandt's (1979) view, according to which a desire is something that is (by definition) causally related to action-tendencies

(but also to certain other things) in a certain way[1].

Now, to regard our desires and aversions as motivational states[2] is (roughly) to regard them as something that (in some sense of the term) "motivate" us to act or to refrain from acting, or as "psychological and physiological promptings[3] to act and refrain from acting" (Kekes (1988), p 18). Or as Griffin (1986) puts it, "[i]n the present technical sense /.../, desiring something is, in the right circumstances, going for it, or not avoiding or being indifferent to getting it" (p 14). That is, the most central feature of desiring is (on this view) regarded as relational; the "essence" of desiring is to be found in how it is related to action[4].

How, then, is desire, *qua* motivational state, related to action? Well, as a first rough approximation, we can say that on the functional conception, every action is a manifestation of some desire, but every desire is not manifested in action. The idea that every action is a manifestation of some desire is basically an idea about how our actions are to be explained: To give a psychological explanation of a particular action consists, in part, in attributing a certain particular desire to the agent[5]. But even though every particular action is a manifestation of some particular desire, there are many particular desires that are not manifested in action. First, our desires may manifest themselves in other ways, e.g., in our thoughts and emotions, and second, it might even be possible that some of our desires are not manifested at all (neither in action nor in thought).

All functional conceptions of desire are (of course) consonant with

---

[1]The reason why I do not think of Brandt's view as dispositional is that he does not regard our desires themselves as behavioural dispositions (or action-tendencies), but as lawfully related to such dispositions (cf. below).

[2]For the time being, I will disregard the idea that desires might be more than just motivational states, i.e. that their functional role may (so to speak) be broader than this, that they may not just be essentially related to action, but also to other psychological states, e.g., to emotion.

[3]But can the latter really be regarded as *desires that*, i.e. as propositional attitudes? I think not.

[4]This suggests that if value-internalism is true, then our positive evaluations must be regarded as desires, and our negative evaluations as aversions.

[5]This idea should be carefully distinguished from two related ideas, viz. (i) the idea that to view a piece of behaviour (or "passivity") *as an action* is, in part, to regard it as *motivated by some desire or other*, and (ii) the idea that to identify (view or describe) a piece of behaviour *as an action of a certain type* sometimes involves attributing a certain desire to the agent ("hunting" or "committing a murder" are examples of actions which are conceptually related to certain desires).

both (iii) and (iv), and it does not seem impossible to construct a functional conception of desire which is consistent with (i) and (ii) as well[6]. This *suggests* that the desire theorist should accept some functional conception of desire, especially if the functional conception is the only alternative to the phenomenological conception.

Now, this is not the place to give a detailed functional account of desire which is consistent with the rudimentary conception above, but I think it is of some importance to say something about what questions such a functional account has to answer, and (also) how these questions might be answered.

The questions that need to be answered if the functional conception is to be given a precise enough formulation can be divided into two groups:

(1) If every particular action is seen as a manifestation of some particular desire, how should the connection between a particular action and the desire which, in part, explains it be characterized? For example, is the relation between the two conceptual or causal, and is it internal or external?

(2) How do we reconcile the fact that desire is paradigmatically manifested in action with the fact that most of our particular desires are not manifested in action? If a particular desire is not manifested in action or behaviour, does this imply that it is not related to action at all? Or is it possible that it is indirectly (rather than directly) related to action? If there is such an indirect relation between unmanifested desire and action, how should it be characterized?

Let us now look at how these questions *might* be answered.

Action as Manifestation of Desire: Desire as a theoretical and explanatory notion

In everyday life, we normally try to make sense of our deliberate actions by giving "common-sense psychological explanations" of them. Now, to give such an explanation of an action is always to explain it in terms of desire, on the one hand, and "belief" (or some other cognitive state or event, like thought), on the other, or more specifically, we explain an agent's actions psychologically by attributing certain desires

---

[6]Conditions (i)-(iv), i.e. my rudimentary conception of desire, can be found in section 4.1, on pp 169-172.

and beliefs to this agent (where the desire and the belief must be appropriately related to one another and to the action they jointly explain). Some philosophers talk about this in terms of theory: They regard desires and beliefs as theoretical constructs which are parts of a common-sense psychological theory, or "folk theory", of deliberate action. Stich (1983) is one good example of such a philosopher, and Brandt (1979) is another: On Brandt's view, "desire" (the term he uses in this context is "*valence*", the genus of which desires and aversions are species) "/.../ is not an observation term, but a theoretical construct in a psychological theory, and its meaning is conferred on it by the laws in which appears /.../" (p 26)[7].

On a very common philosophical view (shared by philosophers as different as Aristotle, Hume and Ryle), this is the way in which action *must* be explained, i.e. *every* psychological explanation (common-sense or not) of action must include a reference to some desire and some belief. Why is this so? Wollheim (1991) is probably right when he claims that desire/belief psychology "is not so much a psychological theory as a pro forma for psychological explanations" (p xxviii). This suggests that desire and belief are internally related to action, i.e. that an action is, by definition, something which can be explained in terms of desire and belief[8]. Or alternatively put, to use the desire/belief schema to explain a piece of behaviour is to *establish it as an action*, since to see a piece of behaviour (or passivity) as an action (i.e. to view it from a psychological perspective) *is* to see it as "explainable" in terms of beliefs and desires[9].

So, it seems that the central issue here is really the following one: For an action to be (successfully) explained by reference to a desire and a belief, (a) how must the desire and the belief be related to one another,

---

[7]This would not just "explain" why actions are explainable in terms of desires and beliefs; it would also explain why we can gain knowledge about what people want (and how much they want it) from observing and interpreting their actions (what they do and do not do). (But notice that we cannot infer what people want (etc.) from action and "inaction" only. For preference to be reflected in choice, certain conditions must be satisfied).

[8]The suggestion is *not*, at this point, that every particular action is internally related to the *particular* desires and beliefs which explain it. Cf. the first idea in note 5 above.

[9]As Wollheim points out (ibid., p xxix ff.), some of Freud's discoveries can be understood in this perspective. For instance, by introducing new explanatory factors (e.g., unconscious desires and beliefs), he "deepened", as Wollheim says, the desire/belief schema. And a part of the "*meaning*" of this deepening was that the number of behaviours we must regard as actions was "greatly enlarged".

and (b) how must they be related to the action they (jointly) explain?

Suffice it to say that as far as (a) is concerned, most philosophers seem to have adopted the idea that a belief is (in this context) appropriately related to a desire if it is *instrumental* vis-à-vis this desire, if it specifies the best way that is, in the circumstances, open to the agent of satisfying the desire[10]. And as far as (b) is concerned, most philosophers have agreed on the following view: For a belief and a desire to be appropriately related to an action (for the purpose of explanation), it is necessary that they make the action *rational* for the agent to do[11]. And since it is (on this view) rational for an agent to perform a certain action only if the agent believes that this very action is the optimal way of satisfying some desire of his, we can see that this idea is almost identical with what was said under (a)[12].

Let us now turn to (2) the question of how the above account of the relation between desire and action can be combined with the fact that most of our desires are not "fully functional". To see how this can (perhaps) be done, let us (for the purpose of illustration) take a closer look at Brandt's (1979) theory of desire.

Desires and Action-tendencies: Brandt's theory

To explain why our desires do not always manifest themselves in action, they must be viewed as connected primarily to action-tendencies or

---

[10]"The best way open to the agent...". That is, the best way open *from the agent's point of view*, rather than the (objectively speaking) best possible way.

[11]But on my view, this does not imply (as e.g., Wollheim, Elster, and Davidson seem to think) that the belief and the desire *constitute* the agent's reason to perform the action. The desire and the belief could also be regarded as giving him, in some way, a reason (as providing him with a reason, or contributing to his reason).

[12]This is where the agreement ends, however. Some philosophers (e.g., Wollheim, Aristotle/Nussbaum, Elster, MacIntyre and Davidson) think that the answer that was just given to (b) is rather insufficient, and they want to add the idea that a desire and a belief cannot explain a certain action unless they jointly *cause* the action. To adopt this view is to conceive of psychological explanation as a kind of causal explanation; to be, in von Wright's (1971) terms, a *causalist* on this issue. Other philosophers (e.g., von Wright, Ryle, and Kenny) have (instead) adopted an *intentionalist* view, on which the connection between an action, on the one hand, and the desire and the belief which explain it, on the other, is conceptual rather than causal, and on which psychological explanations are not causal but *sui generis*. In this context, there is really no need to determine who is right, however; the important thing here is to see what the two positions have in common, viz. that they both conceive of desire as an explanatory notion, i.e. as something which plays a central role in the psychological explanation of an action.

behavioural dispositions, rather than to actual actions and actual behaviour. On this view, to desire something is to have a tendency to act in a certain way, or "to be disposed to" act in a certain way.

So, how is the tendency to perform a certain action related to the actual performance of this action? On Brandt's view, the connection can be characterized as follows: An agent will actually perform action A if and only if "the tendency to perform *A* is stronger than the tendency to perform any other action *B*" (p 26). Brandt himself seems to regard the relation between action-tendency and action as a causal relation, but this is (I think) not necessary[13].

All this leaves room for different conceptions of the relation between desires, on the one hand, and action-tendencies or dispositions, on the other: For example, should our desires *themselves* be regarded as action-tendencies or behavioural dispositions (as Smith seems to suggest), or should they (rather) be regarded as causally or conceptually *related to* (but not identical with) such action-tendencies and dispositions? This is how Brandt conceives of the relation between desires and action-tendencies: To valence (want or be aversive to) a certain situation at a given time *t* is to be in a certain "central motive state". P wants (desires) X (that some situation obtains) at *t* if this central motive state "is such that if it were then to occur to him that a certain act of his then would tend to bring /.../ [X] about[14], his tendency to perform that act would be increased" (ibid., p 26)[15].

---

[13]The same thing holds for the relation between P's disposition to X (e.g., to act in a certain way) in circumstances C and the "episodes" it disposes P to (i.e. X-ing): It too need not be regarded as a causal relation. As Smith (1994) points out, "it is a substantial philosophical thesis to claim that dispositions are causes" (p 114).

[14]If it occurs to me that "if I do A, this is likely to bring X about"; is the occurrence of this idea independent of my desire that X? Probably not, at least not in most cases. As Brandt himself points out: "When some kind of event is valenced there is a disposition to think of such an event, or to notice ways of bringing it about" (ibid., p 27).

[15]The "valence" which Brandt has in mind here is "occurrent valence", but there are also (on Brandt's view) two other fundamental kinds of valence, viz. effective valence and normal valence. This is how Brandt distinguishes between the three kinds of valences: A person P *occurrently wants* (at *t*) a situation X if all that is required at *t*, in order for a relevant change in an action-tendency to occur, is that he have a thought at *t*, to the effect that this action would tend to produce X (cf. ibid., p 28). If such an occurently valenced situation is in fact "before P's mind" at *t*, then this situation is *effectively valenced*. (This suggests that occurrent valences may be viewed as dispositions for effective valences: To have an occurrent desire that X is (roughly) to be disposed to have an effective desire that X in certain

Brandt seems to think that the connection between wants (occurrent as well as effective) and action-tendencies are (so to speak) both causal and conceptual. On the one hand, he explicitly claims that the two are lawfully related to each other, and on a normal interpretation of "law", this suggests that he conceives of the relation as causal. But on the other hand, he tells us (on p 26) that "valence" is a theoretical construct whose *meaning* is, in part, "conferred on it by the laws /.../ relating it to action-tendencies" (my italics), and (again on p 26) that action-tendencies are indicative of valence, and "*partially define its meaning*" (my italics again). And this suggests that the relation is of a conceptual kind. This should probably be understood as follows: Particular desires and particular action-tendencies are causally related to each other, while the type desire (a theoretical construct) is (instead) conceptually related to the type action-tendency.

It should be noted that Brandt's conception of desire (or "theory of valence") includes more than a specification of the relation between valences and action-tendencies. On his view, there are other tendencies (besides action-tendencies) "which are lawfully related to valence, indicative of it, and partially define its meaning" (ibid., p 26). These "tendencies" are of four different kinds, viz.:

(1) Tendencies to be disappointed in certain kinds of circumstances. Disappointment is regarded as evidence of positive valence, i.e. if it makes me disappointed that X does not obtain, then this suggests that I must have had a desire that X.

(2) Tendencies to feel elation and to display "elation behaviour".

(3) Tendencies to think certain kinds of thoughts: "When some kind of event is valenced there is a disposition to think of such an event, or to notice ways of bringing it about" (ibid., p 27). In my opinion, this is the feature of desire which is needed to explain why many of our desires are (like many of the virtues), in Oksenberg Rorty's (1988) terms, *tropic dispositions*, dispositions which lead its "carrier" to

---

conditions, viz. when X (or a part of X) is before P's mind). A person has a *normal valence* for something if he normally (under normal conditions) has an occurrent valence for it. But is it really reasonable to conceive of "normal valences" as "proper desires and aversions"? I think not. What Brandt calls normal valence is better regarded as some kind of second-order dispositions, or "traits (or states) of character". In any case, it doesn't take much reflection to see that "normal valences" are of no interest whatsoever to a desire theorist. Their relevance seems to be mainly of an explanatory kind.

gravitate to certain sorts of situations (cf. p 316). If I desire something I tend to think of it often, and to notice ways in which I can realize it, and this gives in turn (because of the same desire) rise to certain action-tendencies. The occurrence of the cognitions (thoughts, perceptions, etc.) which (when the desire is there) activate the relevant action-tendencies are not always random occurrences. There are (of course) many cases where something presented to me "from the outside" activates ("awakens") my desire, but there are also cases where desires are, so to speak, activated "all by themselves", from the inside.

(4) And at last, desires are also lawfully related to how easy it is to choose: The more "strength" a valence for a certain outcome has compared to valences for other outcomes, the quicker (and the easier) I can (in situations of choice) make a decision[16].

The relation between desire and experience

Before we end this appendix, let us just point out another problem that a functional conception of desire has to deal with. Let us first note that this conception is not really compatible with the common sense idea that we are sometimes directly aware of our desires, or alternatively put, that at least some desires have "felt intensities", or "phenomenological content essentially". (The reason for this is simple: Dispositions are not things of which we can be directly aware)[17]. This means that the functional theorist has to deal with the following question: "If our desires never have experiential content, and if they are never directly experienced, how should the connection between those of our desires that are "seemingly felt" (e.g., bodily appetites like hunger, thirst, or sexual craving) and the relevant conscious experiences (our "feelings of want") be characterized?" Or alternatively put (if we assume that the functional theorist regards the relevant "feelings of want" as representations of desires): "What is it for a (particular) desire to be represented in conscious experience? If a particular desire is represented in a certain particular (introspectible) experience, how should the relation between the

[16]One might ask what notion of strength Brandt has in mind here: is it perhaps motivational force?

[17]However, this does not mean that a functional theorist would deny that "on occasion, when I have a desire, I have certain psychological feelings, analogous of bodily sensations" (Smith (1994), p 105). He does not regard these "feelings of want" as content of the desire, however, but as something with which the desire is linked or correlated, or as something which accompanies the desire.

two be characterized?" (I will not discuss any possible answers to this question, however; I just wanted to present the problem).

This completes the characterization of the functional conception of desire. To repeat, it is likely that the conception of desire which makes the desire theory most relevant is some kind of functional conception, but this is not anything we should take for granted. After all, the important thing is that the desire theorist accepts the rudimentary conception of desire that was presented on pp 169-172 above.

Summary

The main purpose of this thesis is to find the most plausible answers to the following three substantive questions of prudential value (or well-being): (I) What does a person's well-being consist in: what has final value for a person? (II) How do we determine just how valuable a certain situation (or fact) is for a certain person? And (III) how do we determine how well off a person is on the whole (at a certain time)? In section 1.1, I formulate these questions as clearly and precisely as I can.

The way in which I try to achieve the purpose of the thesis is by conducting a critical examination of three common types of answers which have been given to the central questions, viz. hedonistic theories, desire theories, and "objective list theories". (A survey of these traditional answers is offered in section 1.2). For each of these "theories of prudential value" (or "conceptions of well-being"), I first give a formulation of the theory which is as precise as possible (and which makes the theory as plausible as possible). I then try to find out whether the theory in question is a plausible theory, by looking at a number of arguments that can be given for and against the theory. This critical discussion of a number of traditional theories constitutes the major part of the book, and in the course of this discussion, my own theory will slowly take shape.

The first theory I look at is the hedonistic theory. The pure version of this theory can be characterized as follows: (H1) The Experience Requirement: The only facts that can have nonderivative value for a person at a certain time are facts about his or her own experience at that time. (H2) More specifically, the only thing that is nonderivatively good for a person is to have pleasant experiences, and the only thing that is nonderivatively bad for a person is to have unpleasant experiences. (H3) The Thesis of Unrestrictedness: All pleasant experiences are nonderivatively good for the experiencing subject, and all unpleasant experiences are nonderivatively bad, regardless of what other properties these experiences have. (H4) Every good experience is good in virtue of its pleasantness only, and every bad experience is bad in virtue of its unpleasantness only. (H5) The "intensity-orientation" and the idea of proportionality: The value of an experience for the person who has it

is a function of one thing only, viz. how pleasant or unpleasant it is, and this value is (moreover) proportional to how pleasant or unpleasant the experience is. (H6) The final value that a certain life (at a certain time) has for the person who is living it is a function of how much pleasure and how much suffering this life contains. The more pleasure it contains, the better, and the more suffering it contains, the worse.

In chapter 2, I try to formulate these claims in a more precise way. In particular, I discuss how the terms "pleasantness" and "unpleasantness" should be interpreted in this context. What conceptions of the pleasant and the unpleasant do different hedonists have in mind, and (above all) what conception of pleasantness and unpleasantness makes (if combined with the hedonistic theory) the theory most plausible?

In connection with this, I argue that preference-hedonism is more plausible than the quality hedonisms, or alternatively, that the pure version of the hedonistic theory is most plausible if it incorporates the relational theory of pleasantness. On this view, the pleasantness and unpleasantness of a person's experiences are somehow constituted by certain kinds of desires and aversions (likes and dislikes) that the person has. Or more specifically, an experience is pleasant if and only if (and because) the following conditions are met: (i) The experiencing subject has some kind of pro-attitude towards the experience: he desires it, likes it, approves of it, or the like. (ii) The experience is desired (etc.) by the experiencing subject when it occurs. (iii) The experience is desired in a certain way, viz. "in and for itself", i.e. intrinsically, or "as a goal", i.e. "finally". An alternative to (iii) is
(iv) the reason why the experiencing subject desires to have the experience is (at least in part) that it has certain felt qualities.

The main purpose of chapter 3 is to find out, first, whether any version of the hedonistic theory is a plausible theory of prudential value, and second, which version of the theory is most well-founded, pure hedonism or some kind of modified hedonism. After having examined a number of arguments that have (or can) be given for and against different versions of the theory, I reach the following conclusions:

There are a number of strong arguments against pure hedonism, and this theory should therefore be rejected. First, pleasure is not all that matters; there are other things besides pleasure (and experience) which matter to us. How well off a person is (on the whole) is not just depen-

dent on how pleasant his total mental state is, but also on other things, e.g., how much desire-fulfilment there is in his life. Moreover, it seems that certain situations have nonderivative value for a person even though they do not have any pleasant experiential content at all. And second, there are certain pleasures which are not good for us to have, e.g., pleasant emotions the intentional objects of which are "objectively unpleasant".

On the positive side, there is (obviously) some truth in the hedonistic theory. It is not just that it is almost always nonderivatively good for us to feel pleasure; pleasure is also an important good. Moreover, it seems plausible to assume that a person's well-being can not be directly affected (at least not for the better) by things he doesn't know anything about.

I also suggest that the most plausible version of the hedonistic theory is a modified version of the theory, a theory which includes (among other things) the following elements: (R2) If the object of a pleasant emotion is an "objectively unpleasant" situation (e.g., being humiliated), it is not nonderivatively good for the subject to have the emotion. (RW2) It is ceteris paribus better to have pleasant emotions that are based on true beliefs than to have pleasant emotions that are based on false beliefs. The strongest arguments against pure hedonism do not hit this type of modified hedonism, and we should therefore be reluctant to reject it.

I then turn my attention to the satisfaction interpretation of the actual desire theory. The unrestricted version of this theory can be characterized as follows: (D1) Nothing but (actual) desire-fulfilment can be nonderivatively good for a person, and nothing but aversion-fulfilment can be nonderivatively bad for a person. (UD2) The thesis of Unrestrictedness: There are no intrinsic desires that it is not nonderivatively good for a person to have fulfilled, and there are no intrinsic aversions that it is not nonderivatively bad to have fulfilled. (UD3) The positive (or negative) value that a certain desire-fulfilment (or aversion-fulfilment) has for a certain desiring subject is proportional to how strong the desire (or aversion) is. (UD4) The value that a certain life has for the person who is living it is a function of how much desire-fulfilment and how much aversion-fulfilment this life "contains". The more desire-fulfilment and the less aversion-fulfilment a life contains, the

better this life is for the person who lives it.

In chapter 4, I try to give more precise formulations of these claims, viz. by discussing the following four topics:

(1) What is it for someone to desire something? How should the key terms "desire" and "aversion" be understood in this context? What possible conception of desire makes (if combined with the desire theory) the theory most plausible? In answer to this question, I propose that we accept the following rudimentary conception of desire: (i) Desires and aversions are propositional attitudes: the objects of desire are situations (or situations-under-descriptions), and to desire something is to *desire that* some situation obtains, i.e. *that* something is the case, or that some proposition is true. (ii) Desires are pro-attitudes, while aversions are con-attitudes, no matter how the contents (objects) of these attitudes are specified. (iii) Moreover, we should understand the term "desire" in a very broad sense. On the relevant use of "desire", the class of desire includes things as different as volitions and intentions, appetites and longings, projects and purposes, requirements and demands, wishes and regrets, i.e. the class of desire is, in many respects, a very heterogeneous class. (iv) It is also desirable that a conception of desire does not deviate too much from the ordinary uses of the term "desire", especially not from its "explanatory" or "theoretical" use.

(2) What is it for a desire to be stronger than another desire? What notion of strength makes an "intensity-oriented" desire theory most plausible. Here, I suggest that we should understand the term "strength" as rank in a preference ordering (where preference is *not* understood in terms of felt intensity or motivational force). On this view, a person's desire that X is stronger than his desire that Y if and only if he prefers X to Y.

(3) What it is to have a desire fulfilled (or satisfied)? The terms "fulfilled" and "fulfilment" are normally understood in the following way: A person's desire that a situation X obtains is fulfilled if and only if he desires that X and X holds, and a person's aversion to a situation Y is fulfilled if and only if he has an aversion to Y and Y obtains. But does this "traditional" notion of fulfilment really make the desire theory plausible? Is it really plausible to allow for the possibility that a person's well-being can be directly affected by things he does not know anything about, or that it can be nonderivatively good for us to have

our prospective (and retrospective) desires fulfilled? I think not. On my view, we should replace the broad (traditional) notion of fulfilment with the idea that P's desire that X is fulfilled (in the relevant sense) if and only if P desires that X, X holds, *and* the desire and its object are simultaneous (the time of the desire coincides with the time of the occurrence of X). Furthermore, I suggest (in chapter 5) that on the notion of fulfilment which has most "moral and rational significance", a desire is not fulfilled unless the subject is aware of the occurrence of the object.

(4) What is it for a desire to be intrinsic? What notion of intrinsicality makes (if adopted) the "intrinsicality condition" most plausible? Here, I suggest that we should reject both (i) the idea that a situation is desired intrinsically if and only it is desired for its intrinsic properties, or in isolation, rather than for its relational properties, and (ii) the idea that a desire is intrinsic if and only it is underived rather than derived. Instead, we should accept (iii) the idea that a situation is desired intrinsically if and only it is desired "finally", i.e. as an end, rather than instrumentally, i.e. as a means.

The central questions in chapter 5 are questions of plausibility. By looking at a number of arguments that can be given for and against different versions of the actual desire theory, I try to find out first, what possible version of the theory that is the most plausible theory of prudential value, and second, whether this (most plausible) version of the theory a plausible theory of prudential value, i.e. whether any version of the theory is plausible.

This is my answer to the first question: First, the most plausible version of the desire theory is a restricted theory: it claims that only some kinds of intrinsic now-for-now desires should be regarded as relevant. More specifically, if a person P has an intrinsic now-for-now desire that X, the theory claims that it is nonderivatively good for P to have the desire fulfilled (in the traditional sense) if and only if the following conditions are satisfied: (i) X is a part of P's life. (ii) If the desire is derived: It is derivable from the whole truth about its object (and more fundamental intrinsic desires). (iii) If the desire is underived: It is not causally dependent on (maintained by) certain kinds of false beliefs or "ignorances", viz. on beliefs (etc.) whose propositional contents stand in a close enough conceptual relation to the propositional content of the desire. (iv) X is not a situation that is "worth avoiding" (in the pruden-

tial sense). (v) P is aware of the fact that X is (in fact) desired by him. (vi) P is aware of the fact that X obtains. (But with the following proviso: It may sometimes be bad for a person to have an aversion fulfilled, even if he is unaware of the occurrence of its object, viz. if the object has negative prudential desirability-value). Next, the theory also makes certain claims about how we should determine which of two relevant desires that is better for the desiring subject to have fulfilled. The fundamental idea is of course that relevance is a function of strength, but there is one possible exception to this rule, viz. (vii) that desires for situations that are worth desiring (in the prudential sense) are more relevant than desires whose objects are not (in this sense) worth desiring. (It is worth noting that this most plausible version of the desire theory contains certain "objectivist elements", viz. (iv) and (vii)).

So, is this theory a plausible theory of prudential value? Well, its positive claims seem correct: if a person has a desire that is (on the theory) relevant, then it is also good for this person to have the desire fulfilled (in the relevant sense). However, desire-fulfilment is not the only thing that is good for us; it is also good for us to feel pleasure.

The third type of traditional theory is "non-internalist pluralism" (or "the objective list theory". Theories of this type make the following central claims: (1) There are several (universal) prudential values: the facts that have nonderivative value for us are of several types. (2) It is not the case that all the facts that have nonderivative value for a person are internal to this person. (This implies that at least some of the facts that have nonderivative value for a person P are not of the type "P feels pleasure"). (3) At least some of the "non-internal" facts that have nonderivative value for P are not of the type "P has a desire fulfilled".

None of these claims are substantive evaluative claims, however. Non-internalist pluralism is not a substantive evaluative theory, but a type of substantive theory, and the different versions of the "theory" need not ("substantively speaking") have anything in common. This means that no "objective list theory" can really be assessed as such: it is concrete versions of the theory (specific substantive claims about what has prudential value), and nothing else, that can be assessed. So, in order to find out whether the most plausible theory of prudential value is of this type (whether there are any prudential values that fit the description in (1)-(3) above), we need to look at what types of relatio-

nal or external facts non-internalist pluralists have actually regarded as valuable. We can then ask, for each suggested type of fact, whether it is really plausible to attribute prudential value to facts of this type.

In chapter 6, we look at some of the substantive evaluative claims which have been made by various pluralists. The purpose of the chapter is merely to generate a list of possible non-internal facts which have nonderivative value for all human beings. The positive items that seems (to me) most important are classified into seven groups, viz. (1) activities and other "agent-goods", (2) social and relational goods, (3) experiences and other mental states, (4) to be (qua experiencing and thinking subject) in contact with reality, (5) to be a certain kind of person and/or to live one's life in a certain way (to function in a certain way), (6) personal development, and (7) freedom.

In chapter 7, we ask whether any of the claims made by non-internalist pluralists are plausible. A central question here is of course what kinds of arguments that can be given for such claims. Can any universal substantive claims of the form "All non-intrinsic facts of type X are nonderivatively good for all human beings" be justified, and if so, how? In particular, what would an acceptable subject-oriented justification of such a claim look like: what is it about us (about our nature, or "constitution") that makes it nonderivatively good for all of us to have friends, or to be engaged in creative activity?

My conclusion is that we have little or no reason to accept any of the relevant non-internalist claims: In particular, it seems highly unlikely that there is any human nature account that can provide an objectivist subject-oriented justification of the relevant non-internalist claims. And since the other attempts to justify the relevant claims are no good either, and since the counter-arguments against the non-internalist pluralist theories are (on my view) strong enough to place the burden of proof on the pluralists, we should reject all such theories.

Or more specifically, we should reject the idea that there are objective prudential values such that it is good for all of us to "possess" these things, regardless of whether we regard these "objective goods" favourably or unfavourably. That is, we should reject the "tough-minded" (literal, or strong) interpretation of the idea that there are objective and universal prudential values; and we should reject the corresponding strong (pure) version of "the objective list theory".

However, there are "objective prudential values" such that their presence make certain wholes more prudentially valuable than they would otherwise have been. There are "objective prudential values" such that it is ceteris paribus nonderivatively better for a person to take pleasure in these things than to take pleasure in other things, and such that it is ceteris paribus nonderivatively better for a person to have his desires for these things fulfilled than to have his desires for other things fulfilled.

In chapter 8, I present my own mixed theory, a theory which is constructed in such a way so as to be able to stand up to all the objections which have been directed against the other theories. This theory gives the following answers to the central questions (I)-(III):

(I) There are two kinds of situations that are nonderivatively good for a person, viz. (a) to have certain kinds of pleasant experiences, and (b) to have his relevant intrinsic now-for-now desires fulfilled, but only on the assumption that he is aware of the objects of these desires. This answer to (I) is in part hedonistic and in part desire theoretical. "The objective list theory" enters the picture as follows: First, it is not good for a person to take pleasure in something that is on the negative objective list, and second, it is not good for a person to have a desire fulfilled if its object is on the same negative list.

(II) How do we determine just how (nonderivatively) valuable a certain (good) situation is for a certain person? (a) In the case of valuable pleasures, the value that it has for a person to have such an experience is normally a function of how pleasant the experience is. But sometimes, the prudential value of a pleasant experience is also dependent on other things, e.g., on whether it is based on true or false beliefs, or on whether its object is on the positive objective list. (b) In the case of valuable desire-fulfilments, the value that it has for a person to have a relevant desire fulfilled is normally a function of how strong the desire is. But sometimes, the value that it has for a person to have a relevant desire fulfilled does not just depend on how strong it is, but also on other things, viz. on whether or not the object of the desire is worth desiring (in the prudential sense). This part of the answer to (II) is in part hedonistic, in part desire theoretical, in part a combination between hedonism and "the objective list theory", and in part a combination between the desire theory and "the objective list theory". However,

there are also a number of ways in which the hedonistic theory and the desire theory can be combined.

(III) This is how our mixed theory suggest that we determine how well off a certain person is (on the whole, and at a certain time): A person's well-being is (roughly) a function of how much valuable pleasure and how much valuable desire-fulfilment there is in his life. If we formulate the idea in terms of happiness, we get: A person's level of well-being is (roughly) a function of how happy (satisfied) he is with his existence, but only on the assumption that the affective component is based on true beliefs on what his existence is like. But we should also include the idea that there are certain "objective prudential values", viz. in the following way: The happiness (satisfaction) which determines how well off a person is on the whole must (so to speak) "include" how satisfied he is in a number of "objectively pre-determined" areas, and a person's level of satisfaction in the relevant areas must (roughly speaking) be "in line with" the different objective values. In short, to be well off is to be happy for the right reason.

BIBLIOGRAPHY

Abbreviations

NE - Aristotle, *The Nicomachean Ethics*

Allardt, Erik (1993), "Having, Loving, Being: An Alternative to the Swedish Model of Welfare Research", in M. Nussbaum and A. Sen (eds.), *The Quality of Life*, Oxford: Clarendon Paperbacks, 1993

Annas, Julia (1980), "Aristotle on Pleasure and Goodness", in A. Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*, Berkeley and Los Angeles, Cal.: University of California Press, 1980

Aristotle, *The Nicomachean Ethics* (abbreviated *NE*), translated with an introduction by David Ross, Oxford: Oxford University Press, 1991

Beardsley, Monroe C. (1981), *Aesthetics: Problems in the Philosophy of Criticism*, Indianapolis: Hackett, 1981

Bergström, Lars (1990), *Grundbok i värdeteori*, Stockholm: Thales, 1993

Bergström, Lars (1991), "Cykliska preferenser", in W. Rabinowicz (ed.), *Valets vedermödor: sex beslutsteoretiska studier*, Stockholm: Thales, 1991

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

Broad, C. D. (1930), *Five types of ethical theory*, London: Routledge and Kegan Paul, 1930

Broome, John (1993), *The Value of Living*, a manuscript written in 1993

Brülde, Bengt (1992), *Life and Experience*, Göteborg: Filosofiska meddelanden, 1992

Bykvist, Krister (forthcoming), *Changing Preferences*

Chisholm, R. M. and Sosa, E. (1966), "On the Logic of 'Intrinsically Better'", *American Philosophical Quarterly* 3 (1966): 244-249

Cooper, John M. (1980), "Aristotle on Friendship", in A. Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*, Berkeley and Los Angeles, Cal.: University of California Press, 1980

Elster, Jon (1983), *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge: Cambridge University Press, 1987

Foot, Philippa (1985), "Utilitarianism and the Virtues", *Mind*, Vol. XCIV, No. 374 (April 1985): 196-209

Frankena, William K. (1973), *Ethics*, 2nd ed., Englewood Cliffs, New Jersey: Prentice-Hall, 1973

Furberg, Mats (1975), *Allting en trasa? En bok om livets mening*, Lund: Doxa, 1975

Furberg, Mats (1993), *Nedom Vara och Böra?*, Nora: Nya Doxa, 1993

Gauthier, David (1986), *Morals by Agreement*, Oxford: Clarendon Press, 1988

Geach, P. T. (1956), "Good and Evil", *Analysis* 17 (1956-1957): 33-42

Griffin, James (1986), *Well-Being: Its Meaning, Measurement and Moral Importance*, Oxford: Clarendon Paperbacks, 1990

Griffin, James (1996), *Value Judgement. Improving our Ethical Beliefs*, Oxford: Clarendon Press, 1996

Haglund, Björn (1989), "On Act-Utilitarianism", in C. Åberg (ed.), *Cum Grano Salis: Essays dedicated to Dick A.R. Haglund*, Göteborg: Acta Universitatis Gothoburgensis, 1989

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

Hare, R.M. (1981), *Moral Thinking: Its Levels, Method and Point*, Oxford: Clarendon Press, 1991

Harman, Gilbert (1996), "Moral Relativism", in G. Harman and J.J. Thomson, *Moral Relativism and Moral Objectivity*, Oxford: Blackwell, 1996

Harsanyi, John (1982), "Morality and the theory of rational behaviour", in A. K. Sen and B. Williams (eds.), *Utilitarianism and Beyond*, Cambridge: Cambridge University Press, 1982

Hospers, John (1967), *An Introduction to Philosophical Analysis*, 2nd ed., London: Routledge and Kegan Paul, 1973

Hume, David (1739-40), *A Treatise of Human Nature*, Harmondsworth, Middlesex: Penguin Books, 1985

Jeffrey, Richard C. (1983), *The Logic of Decision*, 2nd ed., Chicago: The University of Chicago Press, 1990

Kagan, Shelly (1989), *The Limits of Morality*, Oxford: Oxford University Press, 1989

Kagan, Shelly (1992), "The Limits of Well-Being", in E. F. Paul, F. D. Miller, Jr. and J. Paul (eds.), *The Good Life and the Human Good*, Cambridge: Cambridge University Press, 1992

Kekes, John (1988), *The Examined Life*, University Park, Pennsylvania: The Pennsylvania State University Press, 1992

Kenny, A. (1989), *The Metaphysics of Mind*, Oxford: Oxford University Press paperback, 1992

Korsgaard, Christine (1983), "Two Distinctions in Goodness", *The Philosophical Review*, vol. XCII, No. 2 (April 1983): 169-195

Lear, Jonathan (1990), *Love and its Place in Nature: a philosophical interpretation of Freudian psychoanalysis*, London: Faber, 1990

Liss, Per-Erik (1990), *Health Care Need. Meaning and Measurement*, Linköping: Linköping Studies in Arts and Science, 1990

McDowell, John (1980), "The Role of *Eudaimonia* in Aristotle's Ethics", in A. Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*, Berkeley and Los Angeles, Cal.: University of California Press, 1980

Mill, J. S. (1863), "Utilitarianism", in *Utilitarianism, On Liberty and Considerations on Representative Government*, London: Dent & Sons, 1980

Moore, G. E. (1903), *Principia Ethica*, Cambridge: Cambridge University Press, 1976

Nagel, Thomas (1970), "Death", in *Mortal Questions*, Cambridge: Cambridge University Press, 1979

Nagel, Thomas (1972), "Aristotle on *Eudaimonia*", in A. Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*, Berkeley and Los Angeles, Cal.: University of California Press, 1980

Nagel, Thomas (1986), *The View from Nowhere*, New York: Oxford University Press, 1989

Nordenfelt, Lennart (1991), *Livskvalitet och Hälsa. Teori & kritik*, Falköping: Almqvist & Wiksell, 1991

Nozick, Robert (1974), *Anarchy, State, and Utopia*, New York: Basic Books, 1974

Nozick, Robert (1981), *Philosophical Explanations*, Cambridge, Mass.: Belknap Harvard, 1981

Nozick, Robert (1989), *The Examined Life: Philosophical Meditations*, New York: Touchstone, Simon & Schuster, 1990

Nozick, Robert (1993), *The Nature of Rationality*, Princeton, New Jersey: Princeton University Press, 1993

Nussbaum, Martha C. (1986), *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*, Cambridge: Cambridge University Press, 1989

Nussbaum, Martha C. (1988), "Non-Relative Virtues: An Aristotelian Approach", *Midwest Studies in Philosophy*, XIII (1988): 32-54

Nussbaum, Martha C. (1990), "Aristotelian Social Democracy", in R.B. Douglass, G. Mara, and H. Richardson (eds.), *Liberalism and the Good*, New York: Routledge, 1990

Oksenberg Rorty, Amélie (1988), "Virtues and Their Vicissitudes", in *Mind in Action. Essays in the Philosophy of Mind*, Boston: Beacon Press, 1988

Parfit, Derek (1984), *Reasons and Persons*, New York: Oxford University Press, 1991

Rabinowicz, Wlodek, and Österberg, Jan, "Value Based on Preferences: On Two Interpretations of Preference Utilitarianism", *Economics and Philosophy*, 12 (1996): 1-27

Rachels, James (1986), *The End of Life: Euthanasia and Morality*, Oxford: Oxford University Press, 1990

Russell, Bertrand (1930), *The Conquest of Happiness*, London: Unwin, 1990

Scanlon, Thomas (1993), "Value, Desire, and Quality of Life", in M. Nussbaum and A. Sen (eds.), *The Quality of Life*, Oxford: Clarendon Paperbacks, 1993

Sen, Amartya (1985), "Well-being and Freedom", *The Journal of Philosophy*, Vol. LXXXII, No. 4, April 1985: 185-203

Sen, Amartya (1987), *The Standard of Living*, Cambridge: Cambridge University Press, 1990

Sen, Amartya (1992), *Inequality Re-examined*, Oxford: Clarendon Press, 1992

Sidgwick, Henry (1907), *The Methods of Ethics*, 7th ed., Indianapolis: Hackett, 1981

Smart, J. J. C. (1973), "An Outline of a System of Utilitarian Ethics", in Smart, J. J. C. and Williams, B., *Utilitarianism: For and Against*, Cambridge: Cambridge University Press, 1973

Smith, Michael (1994), *The Moral Problem*, Oxford: Blackwell, 1996

Stich, S. (1983), *From Folk Psychology to Cognitive Science: The Case against Belief*, Cambridge, Mass.: The MIT Press, 1983

Sumner, L.W. (1996), *Welfare, Happiness, and Ethics*, Oxford: Clarendon Press, 1996

Thomson, Judith Jarvis (1992), "On Some Ways in Which a Thing Can Be Good", in E. F. Paul, F. D. Miller, Jr. and J. Paul (eds.), *The Good Life and the Human Good*, Cambridge: Cambridge University Press, 1992

Thomson, Judith Jarvis (1994), "Goodness and Utilitarianism", *Proceedings and Addresses of the American Philosophical Association*, 67.4: 7-21

Thomson, Judith Jarvis (1996), "Moral Objectivity", in G. Harman and J.J. Thomson, *Moral Relativism and Moral Objectivity*, Oxford: Blackwell, 1996

Tranøy, Knut-Erik (1973), "Mill's Pig and the Quality of Pleasure", in *The Moral Import of Science*, Sigma Forlag 1988

Tännsjö, Torbjörn (1993), *Vårdetik*, andra utökade upplagan, Stockholm: Thales, 1993

Tännsjö, Torbjörn (1995), "In Defence of Theory in Ethics", *Canadian Journal of Philosophy*, Vol. 25, No. 4, December 1995: 571-594

Tännsjö, Torbjörn (forthcoming), *Hedonistic Utilitarianism*, Edinburgh University Press

Ullmann-Margalit, E. and Morgenbesser, S. (1977), "Picking and Choosing", *Social Research* 44 (1977): 757-785

von Wright, G. H. (1963), *The Varieties of Goodness*, London: Routledge and Kegan Paul, 1964

von Wright, G. H. (1971), *Explanation and Understanding*, London: Routledge and Kegan Paul, 1971

von Wright, G.H. (1986), *Vetenskapen och förnuftet: Ett försök till orientering*, Stockholm: Bonnier Fakta, 1986

Wetterström, Thomas (1986), *Towards A Theory of Basic Ethics*, Göteborg: Doxa (Oxford) Ltd, 1986

Wilkes, Kathleen V. (1978), "The Good Man and the Good for Man in Aristotle's Ethics", in A. Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*, Berkeley and Los Angeles, Cal.: University of California Press, 1980

Wollheim, Richard (1984), *The Thread of Life*, Cambridge, Mass.: Harvard University Press, 1984

Wollheim, Richard (1991), *Freud*, 2nd ed., London: Fontana Press, 1991

Index

abnormality 117-18; 146; 350-51; 356; 358
absolute goodness 9; 380; 383; 384n
accomplishment 310-12; 345-46
action (in relation to desire) 453-7
action-guidance 399-400; 403; 406-8; 410; 415
action-tendencies 453; 457-60
activity goodness 382n
adaptive desires (preferences) 200; 208-10; 254-6
addiction objection 216-7; 233
additive wholes 7
additivity 70n; 156n
advantage-friendship 292
aesthetic goodness 382n
agent-neutral vs. agent-relative value 379-80; 392-4; 396
aggregation 217-9
aggregation, problem of 21; 156; 217
aggregation in the synchronic, problem of 21-3; 69; 80
aggregation in the diachronic, problem of 21n
Allardt 330; 334
altruism 408
Annas 144n
anticipatory pleasure 149-50; 152
antisocial preferences 206
appropriate vs. inappropriate emotions 145-7
arete 299n
argument from purity 196
Aristotelian appeals to Human Nature 324; 340
Aristotle 52; 127; 144n; 147n; 166; 290-2; 290n; 294; 294n; 297; 299-301; 299n; 300n; 322-3; 325-6; 325n; 326n; 327n; 340; 342n; 343; 354; 364; 364n; 456; 457n
artificial desires 249; 249n; 252
atemporal boundaries of a life 228
atomism (evaluative) 19; 69; 80-1; 109; 155; 215; 217; 219; 284; 310
atomist arguments 308-9; 364
atomistic and summative conception of subjective duration (time) 441-6
attitude model (of pleasantness) 83; 86
attitudinal notion of preference 173n; 174
attributive goodness 15; 381; 417-8
autonomous desires 40; 208; 210; 254-6
autonomous living (functioning)14; 288; 298; 302; 304; 308; 310; 315-7; 323; 345; 357; 362
autonomous preference-ordering 256
aversion 160
aversion, notion of 170
aversion-fulfilment 35; 36; 161
avoidance of harm 335-6; 346-7
awareness-oriented desire theory 268; 370
axiological individualism 397-8; 402
basic need 328-36; 338; 348
basic need account 322; 327; 338-9
basic need satisfaction 336
Beardsley 381; 417-8
behavioural notion of preference 173n; 174n
benevolence 407-9
benevolent desires 205; 409
benevolent norms 25; 409-10
benevolent third parties 25; 411-3

Bentham 85
Bergström 56-7; 58n; 175n; 432-3; 436n
betterness 387; 414-6; 418-9
betterness-for 11; 399; 420
biological vs. non-biological interests 346
bodiliness 353; 355
bodily appetites 169; 180; 460
bodily features 296-7
bodily feelings 76; 78; 90; 146
bodily sensations 75; 77; 79
Brandt 27n; 72n; 74; 84; 87; 87n; 88n; 101-2; 161-2; 182; 182n; 246-9; 249n; 250n; 255; 453; 454n; 456-9; 458n; 460n
Broad 84; 89-90; 101
Broome 415n
Brülde 78
Bykvist 176n; 179n; 180n; 258n
capabilities (abilities) 355-7
causal dependence on false beliefs or states of ignorance 244; 247; 249-53; 260; 271
causal history (of desires) 237; 249; 253-4; 256
causalism 457n
character 293-4; 298-300
character-friendship 292-3
Charles 342n
Chisholm 414-5
choosing (freely) 138; 302; 309; 364
claims concerning relative weights 34; 143; 202; 219-20; 223; 238-9; 260; 263; 266
claims about human nature 351
cognitive psychotherapy 246-53; 255
cognitivism 56-7; 67
coherentism 58n; 62; 62n; 64-5
comparative evaluations 12
comparative questions of prudential value (well-being) 15-7; 22; 422
conceptions of desire 160; 452
conceptions (theories) of human nature 306; 349-50; 359-61
conceptions of rational desire 237-9
contact with reality, value of 138; 151-2; 294
contributory value 20n; 34n; 363
Cooper 290n; 291-3; 342-3
counteradaptive preferences 208; 255
Davidson 457n
death 129-31; 427
deductivism 64-5
definitional deliberation 311-2
deliberative theory of rational desire 237-8; 244-8; 250; 255
deliberatively rational vs. irrational desires 245; 251
derivative vs. nonderivative value-for 6-7
desirability 391-2
desirability-value 258-9; 268-9; 272
desires about one's own life 224; 230
desires (and aversions) as propositional attitudes 161; 166; 170
desire and time 176
desire for 165-9
desire that 165-9
diachronic atomism 20n
diachronic form 19n
diachronic wholes 17-8
Difference Theses 70; 158
discourse ethics 27n
dispositional conception of desire 453
Duncker 84
duration of experience 440

Edgeworth 442-3
egoism 10n; 394; 408
Elster 27n; 57n; 199-200; 209; 253-5; 256n; 298
emotion 76-7; 143; 146; 148; 151
enjoyment 74; 75n
enjoyment-hedonism 75n
entity-view 383
Epicurus 43; 130-1; 227
episteme 300
equipment goodness 382n
ergon 322-7
ergon argument 322-7
essentially unsatisfied desires 166; 180-1
ethike arete 298; 299n; 300; 300n
eudaimonia 325n; 341
excellences (of intellect and of character) 326
excellences of character 299; 301
excellences of intellect 300-1
existence (life) as a whole 13-7
Existentialism 297
experience 29
experience machine 66; 117; 122-4; 137-9
Experience Requirement 29; 32; 43; 68; 94; 124; 128-9; 135
experience-oriented Success Theory 93-5; 128; 206; 223; 231
experiential lives (as wholes) 73; 75; 79-81
external limits 356
externalism 45; 47
extrinsic desire 183-5
fallibilism 437
felt intensity (of desire) 172; 174-5; 208; 211; 213
felt quality (of experience) 89; 100
filia 291-2; 294; 340-1; 345
final value-for 7
final value-for vs. final value-period 395
final vs. instrumental value-period 380; 390-2
final vs. instrumental desire 183-5
first-order desires (preferences) 212-3; 218; 263-5
first-order evaluations 263
first-order preference orderings 213
first-order situations 213
first-person-perspective 127
flexibility 194; 306
Foot 63n
formal features of lives 18-9; 22
formal theories (accounts, criteria) of prudential value 26; 28; 40; 432-3
formal theories in ethics 27n
foundationalism 58n; 62
Frankena 390
freedom 302; 309
Freud 94n; 456n
fulfilment, notion of 161-2; 176; 179; 267
functioning (e.g., in accordance with the Aristotelian excellences) 296-301; 319-22; 325-6
functional conception of desire 172-3; 452-5; 460-1
Furberg 82n; 306n; 349n
Gauthier 173n; 174-5; 246
Geach 15; 381-2
general value-for-claims 2; 59-62-6
generalism 8; 21; 63; 65; 435
genetic theory of rational desire 237-8; 254-5

genetically rational vs. irrational desires 253; 256
global Success Theory 206; 217; 223; 232-5
global vs. local desires 206; 217-9; 223; 232-5
global vs. local situations (facts) 4; 14
goal of (basic) need 328-31; 335-9
good-for-people-statements 3-5
good-making features (characteristics) 65; 416; 418
goodness-for as a relation between an object and a subject 3
goodness-for-P-at-t. 423
goodness-for-people 3
goodness-from-a-point-of-view 379
Griffin 26; 54; 75n; 101; 103; 124n; 129; 135; 136n; 173-4; 173n; 217-9; 233; 237; 242-4; 242n; 243n; 245n; 246n; 276; 288n; 290-1; 294-6; 298; 298n; 308; 310-3; 311n; 312n; 320; 323; 328n; 332; 333n; 336; 339; 345-8; 347n; 362; 430; 454
Griffin's human nature account 323; 345
Griffin's informed desire theory 242
H-evidence 115-7
H-intuitions 115-6
Haglund 401n; 420n; 425n; 427n
happiness 77; 77n; 376-7
HAPPY 141-2
Hare 83; 87; 96; 98n; 167; 176n; 179n; 231; 433n
Harman 57n
Harsanyi 193; 206; 246
hedone 299n
hedonic goodness 382n
hedonic qualities (hedonic tone) 84-6; 89; 91-2; 100-4
hedonic-quality theories 83
hedonism and time 440
heteronomous desires (preferences) 255-6
hexis 299-300; 299n; 300n
higher-order desires and/or evaluations 203; 212-3; 218; 222; 260-5
holism (evaluative) 19
holistic features (of lives) 18; 20
holistic conception of subjective duration (time) 442; 445
Hospers 379
human nature 332; 348
human nature accounts 319; 322-3; 351
Hume 116; 207n; 239-42; 245n; 248n; 256n; 456
Hume's theory of rational desire 237-9
hypothetical choice 174
hypothetical desires (preferences) 273; 275-6
hypothetical persons 277
idealized desire theories 40; 187; 192; 201; 275-8
immanent mistake 281
immanent perspective on value 280-1
impersonal values 11; 397
individuation, problem of 81; 158-9
inductivism 64-5
inferential knowledge (of our desires) 269
informed vs. uninformed desire 237; 242-4
informed theory of rational desire 237-8; 242
instrumental desire 183-5
instrumental goodness 324; 382n

instrumental theory of rational action 199
instrumental value 390
intensity-orientation (of the desire theory) 37; 39; 155; 190; 201-3
intensity-orientation (of the hedonistic theory) 31; 69
intentionalism 457n
interest-utilitarianism 25; 25n; 402; 409
interests (embedded in human nature) 323; 346
internal limits 356
internalism 47; 49
interpersonal comparisons 11
interpersonal contexts 205; 208-10
intertemporal comparisons 11-2
intrapersonal measurements (comparisons) 11-2; 23
intrapersonal context 205; 210; 413
intratemporal comparisons 11-2; 23
intrinsic components of "mixed" desires 163
intrinsic facts (about persons) 47; 49
intrinsic features 388
intrinsic nature 388
intrinsic theory of rational desire 238
intrinsic vs. extrinsic desire 183-6
intrinsic vs. extrinsic value-period 380; 388-90
intrinsicality condition 163; 182; 185
intrinsically rational vs. irrational desire 238; 257-9
irrationality-making features (properties) 248-51
Jeffrey 166; 168
Kagan 14n; 45; 47; 48n; 49; 362; 385; 388n; 391; 391n; 430; 437
Kant 27n
Kekes 44; 171; 214; 261; 262n; 264n; 454
Kenny 165n; 171; 457n
knowledge of our desires 269-70
knowledge-oriented modifications of the desire theory 260; 266
knowledge-oriented objections 212
Korsgaard 383; 388; 390-1; 391n
Lear 7n
life, broad vs. narrow notions 224n; 225n
life-externalism 51
limits 355; 357
Liss 328; 331n
"lives" of social wholes 230
lives-at-certain-times (as wholes) 17-21; 422; 426
lives-over-time 422; 426
living (functioning) 295-6
living-oriented conceptions of the human good 297
Lucretius 130
lupe 299n
MacIntyre 457n
makarios 340
malevolent desires 205; 409
malevolent norms (reasons) 402; 409-10
marginal value of pleasure and suffering 31; 69
marginal value of desire-fulfilment and aversion-fulfilment 37
marginal value of other things 55
masochistic pleasure 145-6
maximally sanctioned by first-order evaluations 263
maximally sanctioned by higher-order desires 263
McDowell 324

merit 417
Mill 32; 85; 92n; 108n; 113-4; 116-7; 121; 133n; 140-1; 141n; 364
Mill's pig 124; 140
minimally sanctioned by first-order evaluations 263
minimally sanctioned by higher-order desires 263
mixed theory 367-76
moderate naturalism 360
modifications of the unrestricted actual desire theory 219-222
modified hedonism 32-5; 102; 105; 124; 135; 139-40; 142; 152; 368
monist (vs. pluralist) theories of prudential value 45; 52
Monistic Quality Hedonism 91; 101
Monistic Quality Theory (of pleasantness) 84-5; 100-1
moods 77-9; 81
Moore 20; 20n; 108n; 113-4; 113n
moral goodness 382n
Morgenbesser 174n
Motivation Theses 403-5
motivational force (of desires) 172-5; 211
motivational theory (of pleasantness) 83; 86
multiplication, principle of 440; 444-5
Nagel 49; 49n; 51-2; 52n; 126-7; 133; 258; 324; 326n; 379; 384; 392-3
Nagel's contented infant 132
Narrow Hedonism 91; 100
need satisfaction 331; 337
non-biological interests 346
non-comparative questions of prudential value (well-being) 15-7
non-hedonic qualities (of experience) 85; 89-90
non-inferential knowledge (of what we want) 269
non-intensity-oriented desire theories 39; 202; 220-1
non-internalist pluralism 45; 53; 56; 286; 287-8; 303
non-perfectionist human nature accounts 319; 322
non-subject-oriented arguments 308; 313
Nordenfelt 157n; 367; 377; 377n
normal vs. abnormal (people) 115-7; 350-1; 356; 358
normative contexts 200; 411
normative importance 416
normative intuitions (judgements) 197-8
normative relevance (of value-for) 9; 24; 379-80; 386; 389; 391; 399-400; 402-3; 410-3; 415
now-for-now desires 95; 180
now-for-nowish desires 179-80
now-for-then desires 162; 176
Nozick 18; 18n; 27n; 57n; 74; 74n; 75n; 77; 77n; 89-90; 106n; 120; 124; 137-40; 146; 149; 151; 294n; 295n; 302n; 304; 342n; 362n; 364
Nussbaum 166; 209-10; 246n; 290n; 291; 294; 300n; 340-3; 342n; 353-6; 353n; 355n
Nussbaum's conception of human nature 352; 357
object interpretation (of the desire theory) 27-8; 40-2; 46; 51; 93-5; 113; 126; 129; 134; 187; 192; 195-6; 274-5; 279; 313-5; 430

object of desire 165
object of need 329
object-oriented (content-oriented) modifications of the unrestricted desire theory 221-3
object-oriented objections to the unrestricted desire theory 203; 205
object-oriented restriction claims 220; 223
object-oriented vs. subject-oriented justification 111
objective duration 441; 444
objective justification 57; 106; 308
objective list theory 29; 44; 52; 53; 286; 303; 428-31; 437
objective (prudential) values 135; 362-4; 377; 434; 438
objective reasons (for acting) 25; 107; 384; 392-3; 410
objective value 258; 390; 392
objectively pleasant vs. unpleasant 145-6
objectivism 428; 434-6
objectivist justification 314; 317; 319; 322
Oksenberg Rorty 459
ordinary language (speech, use, meaning) 72-5; 90; 96; 99; 102; 103; 160; 166; 171
orexis 166
organic unity 18; 19; 22
other-regarding desires 206; 228-9
pain 82-3
Parfit 8n; 29; 44-6; 74; 87; 91; 93; 94n; 100; 109-10; 109n; 139; 152; 182; 196n; 201n; 205; 206n; 212; 216-7; 217n; 219; 226; 229; 232-3; 237; 246; 257; 257n; 258n; 260; 278; 282n; 286; 288; 317; 367; 376; 428-9
part of a person's life 225
partial justification 59-60; 65; 67
partial view 79-81
particular value-for-claims 60; 65; 67
particularism 21
pathos 299n
perfectionist account of good human functioning 320
perfectionist human nature accounts (justifications) 319; 322
person (human being) 4n
person-externalism 50
person-oriented conceptions of the human good 297
personal development 301
personal values 397
phenomenological conceptions of desire 172; 173; 213; 452-3; 455
phronesis 298; 300-1; 300n; 325
pleasure-friendship 292
Pluralist Quality Theory (of pleasantness) 85-6; 101-3
pluralist theories of value-for 45
Pluralistic Quality Hedonism 92; 103
post-mortem events 127; 226
potentialities 297; 302
praxis 299n
predicative goodness 380-1
Preference Autonomy, principle of 193
preference change by framing 254
preference orderings 172; 175; 206; 208; 211; 213; 218; 255; 265
preference-egoism 199
Preference-Hedonism 93-5; 103; 232
preference-theory of (pleasantness) 83; 86
preference-utilitarianism 199; 209

pro-attitudes 170
proportionality (between values and other things) 23; 31-2; 37; 54-5; 69; 155
propositional attitudes 167; 170
propositional content (of desires) 161; 163; 167; 179
prospective desires 162; 176-7; 179-80
psuche 325
psychological hedonism 113-4; 117; 119; 352
pure hedonism 29-32; 68-9; 71; 73; 91; 105; 110; 124; 132; 139; 140; 142; 150; 152
pure (simple, extreme, or raw) non-internalist pluralist theories 54; 304; 362; 364
qualitative hedonism 85
Quality Hedonism 91-2
Quality Theories (of pleasantness) 83; 85; 96; 98-9
quality-of-experience theories (of pleasantness) 83
quantitative hedonism 85
quasi-subjectivist justification 313; 315-6
questions of prudential value (well-being) 1-3; 15; 22; 24-5; 407; 410; 412-3; 415
Rabinowicz 27-8; 28n; 182; 187n; 274n; 281; 396n
Rachels 136; 360
radical deliberation 311
Ratio Theses 71; 158
rationality-oriented (and awareness-oriented) Success Theory 272; 284
rationality-oriented modifications of the desire theory 220-2; 236; 238-60
rationality-oriented objections to the unrestricted desire theory 203; 207
rationally required desire 257
Rawls 338
recognizably human lives 323; 341; 343
reflective equilibrium 62-3
relational goods 290
relational theory (of pleasantness) 83; 86; 89-90; 96-7; 99; 102-3
relationalism 45; 47; 50
relativism 5; 46; 316; 436
relevant vs. irrelevant desires and aversions 38
restricted hedonism 32; 34
restricted versions of the desire theory 37; 196
restriction 219-21
restriction claims 34; 38; 143-4; 224; 238; 259
retrospective desires 162; 176; 178-9
Ross 340; 381
Ruse 350
Russell 20; 20n
Ryle 457n
sadistic pleasure 145-6
sadistic desire 199
satisfaction (with life) 376-7
satisfaction interpretation (of the desire theory) 27-9; 35; 40; 42; 45; 51-2; 154; 187; 192; 195; 274-5
Scanlon 26; 42; 193; 193n; 197; 220n; 242n; 279-80; 282n; 283-4; 283n; 411-3; 429-30; 433n; 436-7
Scanlon's perspectives 411
second-order desire 264
self-interest 412
self-interest theory of rationality 198; 402; 408

self-interested norms 25; 410
semantic content (of desire) 167
semi-general value-for-claims 64; 65
Sen 209-11; 281; 283; 283n; 285n; 302; 309-10; 309n
sensation 76; 82; 97
sensation models (of pleasantness) 83
sensory component thesis 99
Sidgwick 88; 88n
Significant Others 228-9
situation (fact) 4n
situation-view (of the object of desire) 165-8
situation-view (of carriers of value) 383-4
Smart 108n; 141
Smith 172; 452-3; 458; 458n; 460n
social choice theory 200; 205
social justice 199
social wholes 230-1
social-utility function 206
Socrates 133n; 140; 140n
sophia 299-301; 325
Sosa 414-5
Sovereign (Desiring) Subject 193; 264
spirituality 361
spiritual prudential values 361
standards of (e.g. prudential) goodness 15; 417; 421
Stich 456
strength (of desires and aversions) 161; 172-5; 206; 208; 210; 213; 217-9; 235; 265
strong phenomenological conception of desire 452
subject-oriented arguments (for and against different theories of prudential value) 112; 119-20; 191-2; 307; 313-4; 316
subject-oriented vs. object-oriented justification of value-for-claims 110-2; 421
subjective duration (e.g. of experiences) 440-2; 444-6; 448
subjective evaluations 115
subjective justification 57-8; 66
subjective reasons for acting 107; 410
subjective value 379
subjectivism 8; 46; 428; 434
subjectivist justification 313-4
substantive vs. formal theories of prudential value 26-8
substantive good theories 26; 45; 429-30; 437
Success Theory 206; 223-4
suffering 82-3; 87; 96; 98
summative theories 216; 234
Sumner 9n; 26n; 173n; 176n; 177n; 193-4; 193n; 197n; 233n; 261; 267n; 288; 288n; 306-7; 306n; 307n; 315; 317; 327n; 335; 367; 377; 377n; 433n
supervenience 15; 18; 65; 399; 416; 418-21
synchronic wholes 17-9; 21
synchronic atomism vs. synchronic holism 20-1; 23
synchronism 179-81; 179n; 186
Taoism 297
techne 300-1
technical goodness 324; 382n
teleological need 328-30
temporal boundaries of lives 226
temporal perspectives 442; 446
tension-need 328-9
theoretical ambition in ethics 63
theoretical rationality 57n

theoria 301
theory of simultaneous satisfaction of preferences 179n; 181
thesis of Additivity 233
Theses of Unrestrictedness 30; 36; 68; 155; 189
thing-view (of object of desire) 165-6
third-person-perspective 127
Thomson 3; 3n; 5-6; 6n; 63n; 379; 381-2; 382n; 383n; 384n; 387-8; 390; 400; 403-4; 404n; 407; 414
to-be-avoidedness 387
to-be-preferredness 387
to-be-pursuedness 387; 389; 391-2; 395
Tolland 391n
total view 79-80
transcendent perspective on values 281-2; 281n
Tranøy 140
tropic dispositions 459
trumping (lexical priority) 54
Tännsjö 58n; 62-3; 179n; 196n; 261; 277; 278n; 279n; 298; 442-3; 442n
Ullman-Margalit 174n
underived (vs. derived) desires 182-3; 185
uninformed desire 242; 244
universal needs 331; 333; 336
universal prudential values 1; 9; 59; 286; 304; 307; 310-2; 315-6; 438-9
universalism 316; 435-6
universalist pluralism 46
unrestricted (actual) desire theory 36; 154; 189
utilitarianism 9; 10n; 64n; 402; 409; 412
utility 173n; 174; 174n
value as a matter of degree 12
value-for 3-13; 394-8
value-for-internalism 404-5
value-for-P-at-t 226-7
value-period 380-94
value-period-internalism 403-4
Velleman 7n
virtue 298-9
virtue-friendship 292-3
von Wright 5; 324; 335n; 382n; 457n
Wants Theses 403-5
weak cognitivism 56; 59
Weak Wants Theses 403-5
weak phenomenological conception of desire 452
weaker versions of the objective list theory 362
well-being, notion of 13
well-being and time 16; 422-7
well-functioning person 26; 146; 150-1; 301; 325-6; 364
Wetterström 12; 115-8; 116n; 118n; 383-4; 388n
Wilkes 324-5; 326n
Wittgenstein 101n
Wollheim 225n; 456; 456n; 457n
Wonmug 136-7; 151
Zen 297
Österberg 27-8; 174n; 182; 186n; 187n; 281; 396n

ACTA PHILOSOPHICA GOTHOBURGENSIA
ISSN 0283-2380

Editors: Mats Furberg, Per Lindström, and Torbjörn Tännsjö

Published by the Department of Philosophy of the University of Göteborg

Subscription to the series and orders for single volumes should be addressed to:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222, SE-405 30 Göteborg, Sweden

VOLUMES PUBLISHED

1. MATS FURBERG, THOMAS WETTERSTRÖM and CLAES ÅBERG (editors): Logic and Abstraction. Essays dedicated to Per Lindström on his fiftieth birthday. 1986. 347 pp.

2. STAFFAN CARLSHAMRE: Language and Time. An Attempt to Arrest the Thought of Jacques Derrida. 1986. 253 pp.

3. CLAES ÅBERG (editor): Cum Grano Salis. Essays dedicated to Dick A. R. Haglund. 1989. 263 pp.

4. ANDERS TOLLAND: Epistemological Relativism and Relativistic Epistemology. Richard Rorty and the possibility of a Philosophical Theory of Knowledge. 1991. 156 pp.

5. CLAES STRANNEGÅRD: Arithmetical realizations of modal formulas. 1997. 100 pp.

6. BENGT BRÜLDE: The Human Good. 1998. 490 pp.

ISBN 91-7346-324-8

Kompendiet, Göteborg
1998