The prisms of moral personhood

Schriften der Luther-Agricola-Gesellschaft 33

JAANA HALLAMAA

The prisms of moral personhood

The concept of a person in contemporary Anglo-American ethics

Luther-Agricola-Society
Helsinki 1994

Luther-Agricola-Seura
PL 33 (Aleksanterinkatu 7)
FIN-00014 Helsingin yliopisto
Finland

ISBN 951-9047-37-9
ISSN 1236-9675

Vammalan Kirjapaino Oy
Vammala

Abstract

DESPITE the popularity of moral philosophy and ethical literature it is not clear whether any joint issues belie the multiplicity of ethical approaches. Still, ethical theories are usually philosophical explications of the moral language people use and of the institution of morality. These explications cannot be made without referring to human beings as moral persons. In this sense, "person" is one of the most central conceptions in any moral theory irrespective of its form. The present study takes the centrality of the concept of a person as its starting point and proceeds to ask how different kinds of moral theories imply the concept of a person, which kinds of concept of a person are included in such theories and finally, what significance the concept has for understanding the nature of moral theories. These questions are explored by using models of three different types of ethical theories — utilitarian, contractarian and virtue theories — as material for the analysis. The analysis reveals three different types of a "moral person". Theories which belong to the same group on the basis of the form of the normative moral theory represent a similar concept of a person which, again, differs from the notions of "person" provided by the other two groups of theories.

The theories of Richard Brandt, Richard M. Hare and Derek Parfit form the basis for the examination of the utilitarian theories in the study. The utilitarian concept of a person is a desiring agent who aims to realize her preferences by her action. The role of morality is to qualify the way in which the person naturally acts but moral and non-moral action do not basically deviate from each other.

The contractarian theories which the study examines are by John Rawls, David Gauthier and Alan Gewirth. These theories define the constituents of moral personhood through the sufficient and necessary conditions of intentional action. The same conditions also explicate the

foundation of the institution of morality: it is based on a contract guaranteeing everyone the conditions by virtue of which their intentional agency becomes possible.

Thirdly, the study examines two schools of thinking in modern virtue ethics, the first including the theories of Philippa Foot and James Wallace and the second those of Martha Nussbaum, Alasdair MacIntyre and Charles Taylor. The notion which emerges from these theories can best be characterized as a narrative conception of a person: a person is the narrative of her life, something that is determined by historical and cultural contingency.

The last part of the study analyses the significance of the concept of a person as far as moral theories are concerned. The definition of the morally relevant and the notion of a moral person are closely connected. Moral theories are models for a good life and they are always designed for a certain kind of a person. In this sense, every moral theory incorporates a conception of a person within its premises. This conception is not ethically neutral but implies a view of what is good, normal or meaningful in and for a human life. This result gives rise to a conclusion that has a practical bearing: the concept of a person cannot be used as a basis for solving ethical disputes, for each version of the concept is likely to imply positions that are already normatively loaded.

Acknowledgements

As persons we are never entirely of our own making and the same applies to all our projects; we could not have accomplished them just by ourselves. As I relinquish the writing of this study I am deeply aware of the fact that it would not exist had I not had the privilege of so many people's help and advice.

During the past five years spent, more or less intensively, struggling with the concept of a person, I have been working at the Department of Systematic Theology at the University of Helsinki. It is not easy to think of a better place to work: not only has the department offered an intellectually stimulating place to study, but it has also been a community of magnificent people. Professor Simo Knuuttila, the supervisor of my work, has never during these years ceased to surprise me by his extensive and profound knowledge of all fields of philosophy and theology. His unfailing encouragement, uncompromizing criticism and perpetual support have formed an unbeatable combination of tutorial skills. He has been the best teacher imaginable and I owe him a deep but pleasant debt of gratitude. Professor Heikki Kirjavainen is the person to whom I owe my "professional" interest in philosophy. From the very first years as an undergraduate student until the present day I have constantly profited from his philosophical curiosity, his analytical sharpness and his intellectual honesty. The years which I spent as his assistant are among the best of my life. A well-deserved thanks belongs to professor Reijo Työrinoja, as well. Had it not been for him the many hours I have spent in the coffee room of the department discussing philosophy would have been deplorably wasted; I have learned much more by listening to him than by reading many a book. I also wish to direct a word of thanks to the chair of our department, professor Tuomo Mannermaa who is largely responsible for making the department such a splendid place to work. A special

word of gratitude belongs to Eeva Partio, who has never over these years failed to offer help, encouragement and support. There are many other colleagues whom I cannot thank by name here but who are still no less important as contributors to this work.

At the last stage of this study I also had the privilege to work with professor Brenda Almond, who was kind enough to read and comment on the manuscript as well as to act as my examiner. This opportunity for co-operation has been a source of both intellectual pleasure and personal joy. I thank her most warmly. The last summer that I spent struggling with this work also acquainted me with lecturer John Calton who kindly helped me through the various difficulties of English language. The sessions we have had discussing not only English grammar and the concept of a person but various other things as well have been a welcome oasis of entertainment in the middle of a busy summer. I am both glad and thankful for the fact that this study is published as a volume of the publications of the Luther-Agricola-Society which has been an integral part of my life for so many years. My joy is even greater because Eliisa Isoniemi has designed the cover of this book and research fellow Olli Hallamaa has made the layout for it. Diakonisches Werk has assisted me by granting a scholarship at the initial stage of this work and Oskar Öflunds Stiftelse has offered me financial support in the final stage for which both these institutions deserve sincere thanks.

The people to whom I remain in deepest gratitude but to whom I am least able to express it are the members of my family. I would have interrupted this slow and hopeless process long ago had I not learned at home that perseverance and patience are the best methods for overcoming difficulties. The home of my childhood is the basis from which this work rises. I thank my parents Pirkko and Lauri Pyrhönen for their constant support and the interest they have shown in this study during the years. As far as this work is concerned, my sister Heta Pyrhönen has a special position even among the members of my immediate family. The weekly discussions we have had about various subjects connected to the theme of this work and to the process from which it has emerged have meant very much to me. Those who best know the meaning of despair

and agony in respect of this work are my husband Olli and our children Tuomas and Inkeri. Without the love, sharing and the mutual support that life with them has been this work would not be, nor would I. It is good to bring this study to a completion knowing that there will soon be a member in our family who has not had to live under the shadow of the concept of a person.

I dedicate this book to Olli who is my best friend.

Helsinki 19.9.1994

Jaana Hallamaa

Contents

Introduction

ETHICS has become a popular subject, a field of philosophical study which captures the interest of both experts and "ordinary" people. But what is the subject matter of ethics? We can identify such questions as "How should one live?", "What is my duty?" and "What is the best thing to do?" as ethical questions, but it is difficult to determine what actually makes them so. No simple answer is found in ethical literature. On the contrary, one may ask whether any joint issues belie the multiplicity of approaches or whether different versions of moral philosophy simply represent a variety of mutually incompatible habits of thought.

Independently of what else ethical theories may be, they are usually philosophical explications of the moral language people actually use. By formulating such explications, moral theorists try to clarify what is at issue in people's moral deliberation and what is the best way to understand the nature of normative moral rules, principles and the like that people apply in their everyday reasoning. Such explications cannot be made without referring to human beings, and more specifically, to human beings from a *moral* point of view. Ethical theories could, thus, be characterized as attempts to express what it is to live as a human being in a moral realm, or in other words, what it is to be a *moral person*. Presumably, all moral theories include a concept of a person. This, in turn suggests that we could clarify the nature of ethical theories by studying the concept of a person in them. Before this idea can be developed further, though, it is necessary to make some clarifications and distinctions.

The terms *morality* and *moral* are used for various purposes. The two opposites of the adjective reveal two of these uses: some things, deeds or the like are called *immoral*, whereas other objects are referred to as *non-moral*. As an opposite to *immoral*, *moral* means "right" or "good". In this use, morality is a normative system of principles, rules, etc., the purpose

of which is to direct people's behaviour.[1] Contrasted with non-moral, moral signifies what belongs to the field of morality as distinct to, say, aesthetics or psychology. This meaning of moral squares with "ethical", and we can speak of ethics as the non-normative study of what is moral. Related to this use of the term, we also speak of morality when we refer to the specific institution of morality.[2]

There are also what we call moral theories or ethical theories. We can examine these theories from two different perspectives. First, they offer philosophical explanations of morality in the form of *theoretical* models for understanding what is involved in this human institution. Second, moral theories usually include a *normative* element in the sense that the theoretical model for understanding the nature of morality is developed into an auxiliary for moral reasoning. As theoretical models, moral theories describe what morality is, what makes something part of morality, or what is morally relevant. They define the criteria for regarding something as an ethical question.[3] As a normative auxiliary, a moral theory helps people recognize situations that are morally significant and enables them to see their moral decision-making process in terms of the theoretical scheme the theory offers.

It is assumed in this study that the concept of a person is of relevance from both these points of view. Thus, a theoretical definition of the distinctly moral is somehow connected with a description of *moral* personhood. Similarly, as a normative device, a moral theory involves some

---

[1] Williams gives only this meaning to "morality": "[...] morality should be understood as a particular development of the ethical, one that has a special significance in modern Western culture. It peculiarly emphasizes certain ethical notions rather than others, developing in particular a special notion of obligation, and it has some peculiar presuppositions." Williams 1987, 6. For a similar understanding of the term, see Nussbaum 1990, 169.

[2] Frankena 1973, 5–6.

[3] Moral theories as theoretical models of morality also include questions which have been called *meta-ethical*; see Frankena 1973, 5. It is, however, uncertain whether meta-ethics and normative ethics can be distinguished as clearly as it has sometimes been suggested. These two are rather linked together and, thus, it is meaningful to approach ethical questions from a point of view relevant for both meta-ethical, or theoretical study and normative morality; see von Wright 1968, 6.

explication of how one should deliberate and act as a moral person. According to this assumption, the concept "person" occupies a central position as a theoretical and as a normative notion in any moral theory in the sense that there is some connection between the theoretical definition of the morally relevant, the model for moral reasoning and the concept of a person. We can establish this link between the concept of a person and the basic theoretical and normative formulations of any moral theory if we can show that the concepts of a person explicable in moral theories correspond with the manifest differences between different kinds of moral theories. If our initial assumption is correct we should find, to take an example, a utilitarian concept of a person, common to utilitarian moral theories, but different from the concept of a person which is manifest in contractarian models of moral thinking. Coming to this conclusion would show that "person" is a central moral concept which is closely connected to the way different ethical theories understand the institution of morality. Before the aim of this study can be formulated more specifically we must make some further distinctions concerning the nature of moral theories.

Questions people recognize as moral are very varied and this applies to moral theories as well. A narrow definition of what a moral theory is easily disqualifies approaches otherwise relevant to the theme. I try to solve the problem by seeking the minimum conditions any moral theory must fulfil to qualify as such. This also serves to exclude from the study other approaches to morality, say, anthropological and psychological. A minimum condition of any moral theory is that it should place morality in the context of human *intentionality*; in moral theories intentionality is a necessary condition for something to count as moral. Here we are immediately confronted with two problems: first, mere intentionality is not a sufficient condition for distinguishing morality from other intentional human activity, but there must be further criteria for classification. Second, "moral" is an attribute which can be attached to various things, and not all of them seem to belong to the category of intentionality directly. People can be moral, but so can their deeds: where then can we place morality?

The problem behind the relation between morality and intentional action in general is how morality is related to other considerations concerning reasons for action. There are three possibilities: first, moral reasons for action can be wholly reduced to other reasons for action, e.g., to rationality. In this case, we can cease to use moral language, rather translating it into other modes of speech concerning human intentionality; second, it is conceivable that moral reasons square with other reasons for action, but that it is still meaningful to use moral language as a specific mode of speech; third, morality can be seen to provide an independent and genuine class of reasons for action which cannot be reduced to any other reasons for action without leaving a residue. In this case there has to be a model for clarifying the specific nature of morality and an answer to the question "Why should I express solidarity towards my moral self?"[4]

Another difficulty we face here concerns the place of morality in regard to intentionality. To clarify the issue, let us consider a classical model for analyzing intentionality, the *practical syllogism*:

A wants that p.
A believes that not p if not q.
A starts doing q.

Here an agent's want concerning an end and beliefs regarding appropriate means for achieving that end, form the basis of action. Correspondingly, there has to be a model for explaining the moral relevance of wants and beliefs and of their mutual relation as the spring of intentional

[4] We can think of a fourth possibility concerning the relation between morality and intentionality: instead of looking for reasons for moral action we can seek the causes of such action. According to this alternative, there is no independent source of moral action, but it can simply be reduced to other types of human activity, arising from, say, attempts to fulfil desire or the satisfaction of biologically and psychologically determined needs. Sociobiological theories offer an example of this way of thought; see, e.g., Robert J. McShea, *Morality and Human Nature: a New Route to Ethical Theory*. Temple University Press, Philadelphia, 1990. At its crudest, this approach reduces not only morality but intentionality to other forms of explanation. Accordingly, we cannot genuinely speak about intentionality and reasons for action because everything is, in the last instance, fully reducable to physicalist and mechanistic explanations. This would represent a behaviourist approach to morality.

action. Furthermore, there has to be a scheme for determining the moral role of the context in which these actions take place, e.g., a social community. We can make enquiries as to the moral relevance of the desire and belief attitudes of a person, the moral relevance of an action and of a context.

There are two main points of view to a person, which are relevant in a moral context. A person can be either a moral subject or a moral object. As a moral subject a person is an initiator of something that has moral relevance, usually of deliberative action. In this case, moral relevance can be found in the constituents of deliberation: in wants, desires, beliefs, knowledge and/or the like. Moral significance can residue in a person as an object of moral concern. More specifically, there can be something in the person or in her[5] situation which makes her a *moral* object. In this position a person is not necessarily an intentional agent.[6]

Action, too, can be the locus of morality. There are two ways in which morality might be located within intentional action: actions can be classified according to an agent's intention or according to the external effects of an action.

The third alternative source for the morally relevant is the context in which intentionality takes place. Here we can distinguish between a narrow and a wide context. A narrow context is the particular situation in which intentional activity or action takes place. Morality belongs in a wide context where moral significance is determined by a socially consti-

---

[5] This is a theoretical study concerning the concept of a moral person which makes no explicit reference to questions of gender. Gender may, however, have an impact on the way we understand the concept even in this theoretical context. To keep in mind that the aspect of gender may have relevance, I use the feminine form of the third person singular pronoun to refer to a person in general. The odd impression this sometimes gives may indicate that gender is a relevant issue for the theme although it receives no explicit attention in this study.

[6] To take an example, a situation becomes moral if the interests of other people are concerned. Here the fact that another person has interests makes her a moral object and this is a sufficient reason for regarding the situation as morally relevant. A moral object does not necessarily have to be a human person, but some moral theories also classify (higher) sentient beings as moral objects. There are theories which allow the moral subject and the moral object to coincide, which means that the theories acknowledge duties towards oneself.

tuted morality or by social practices and mores.

These are the theoretical alternatives for locating the morally relevant within a moral theory. The task is now to explore 1) how the theoretical and the normative conceptions of various moral theories imply the concept of a person; 2) which kinds of concepts of a person are included in the different moral theories, and 3) what significance the concept has for understanding the nature of moral theories.

As has often been noted, Anglo-American moral philosophy has undergone a change during the past 30 years. But before that there was a strong tendency to ignore substantial moral questions and to concentrate on what was called *metaethics*. The idea behind the approach was to reveal the nature of moral questions by conceptual analysis, which would then help solve moral problems. The study of moral language did not, however, cover all issues of relevance to the subject, and since the 1970's ethicists have revived an interest in moral philosophy which also deals with more practical ethical issues. There is now a rapidly growing variety of applied ethics whose purpose is to deal with moral problems in specific areas of life.[7] The present study concentrates, however, on works which do not strictly belong to the field of applied ethics but are more general expositions of morality.

As part of this revived interest in substantive ethical issues the concept of a person has received due attention. We can distinguish two main rallying points in the debate. First, there are writers who focus on personal identity.[8] The second major issue concerns the role of the concept of a person in applied ethics.[9] Both topics have an impact on ethical theories and accordingly I relate to both viewpoints in the study. As distinct from these approaches, I am interested in the *concept* of a moral person,

---

[7] This shift can be traced to the disappointment over the project of analytical ethics: notwithstanding its explicit programme it could not transform moral dilemmas into conceptual problems, which could then have been solved by philosophical analysis. This shift of interest is clear, for example, in Richard M. Hare's work. He started out as a strictly analytical ethicist (*The Language of Morals*, 1952) but although his basic ideas have remained unaltered through the years his later application of the basic ideas of his *universal prescriptivism* has undergone an essential change: the ideas are applied to normative questions (see *Freedom and Reason*, 1963 and *Moral Thinking*, 1981).

that is, in the varieties of the role of the concept in ethical theories.

The material for this study consists of different types of recent Anglo-American moral theories from 1952 to 1990. Works by 11 moral philosophers form the basic texts for a detailed analysis, and writings of other ethicists are used as reference material. The aim is not to present a historical review of the development of the Anglo-American ethical discussion during this time. Instead, the chosen theories are treated as representative examples of different types of approaches in modern moral philosophy. They are models by well-known authors whose work has evoked a lively debate, writers who are known to have developed an original moral thinking, or whose work establishes an innovative perspective for moral philosophy.[10] As there is no classification in terms of the concept of a person, the study follows the traditional division of normative moral theories respecting the authors' own classification. Accordingly, the study deals with three groups of theories. After dealing with utilitarian theories,[11] the study moves on to contractarian models.[12] The third part then concentrates on theories of virtue.[13] Each part concludes with a discussion of the extent to which a mutual concept of a person is involved in those theories. The last chapter of the work presents the

---

[8] There has been a lively discussion on the topic of personal identity since the 1970's, see, *The Identities of Persons* (ed. by Amelie RORTY, Berkeley, Los Angeles and London 1976) and Bernard WILLIAMS, *Problems of the Self*, (Cambridge Univeristy press, Cambridge, 1973). The discussion has been spurred on by the ideas of Derek Parfit's book *Reasons and Persons*, (Oxford 1984). Within the identity discussion one can further distinguish a line focusing on the possible models for defining a concept of identity to fit the scientific knowledge concerning the neurological structure of the human brain. On the other hand the discussion centres around such topics as the continuity of one's identity, change and permanence.

[9] See, e.g., *Human Beings*, ed. by David COCKBURN, Royal Institute of Philosophy Supplement: 29, Cambridge University Press, Cambridge, 1991.

[10] This applies especially to four authors whose theories I will study, namely to Richard M. Hare, John Rawls, Alasdair MacIntyre and Derek Parfit.

[11] I will analyze the theories of three utilitarian thinkers: Richard M. Hare, Richard Brandt and Derek Parfit. The works I will deal with in this section are Richard M. HARE, *The Language of Morals*. Oxfod University Press, Oxford, 1952; *Freedom and Reason*. Oxford University Press, Oxford, 1963; *Moral Thinking: Its Levels, Method and Point*. Clarendon Press, Oxford, 1984; Richard BRANDT, *A Theory of the Good and the Right*. Clarendon Press, Oxford, 1979, and Derek PARFIT, *Reasons and Persons*. Clarendon Press, Oxford, 1984.

results of the analysis and returns to the question posed at the outset of this study can the concept of a person help us understand the nature of ethical theories?

[12] In this section I deal with the theories of John Rawls, David Gauthier and Alan Gewirth. I will concentrate on the following volumes: John RAWLS, *A Theory of Justice*. Clarendon Press, Oxford, 1972; David GAUTHIER, *Morals by Agreement*. Oxford university Press, Oxford, 1986; and Alan GEWIRTH, *Reason and Morality*. The University of Chicago Press, Chicago and London, 1978.

[13] I divide the virtue theories into two groups. Philippa Foot, James Wallace represent an earlier approach to the question concerning the meaning of virtue for morality, whereas Martha Nussbaum, Alasdair MacIntyre and Charles Taylor are introduced as explinents of more developed versions of virtue ethics. The works forming the material of my study are: Philippa FOOT, *Virtues and Vices and Other Essays in Moral Philosophy*. Basil Blackwell, Oxford, 1978; James D. WALLACE, *Virtues and Vices*. Cornell University Press, Ithaca and London, 1978; Martha NUSSBAUM, *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge University Press, Cambridge, 1986; *Love's Knowledge: Essays in Philosophy and Literature*. Oxford University Press, New York and Oxford, 1990; Alasdair MACINTYRE, *After Virtue: A Study in Moral Theory*. 2nd edition. Duckworth, London, 1987; *Whose Justice? Which Rationality?* Duckworth, London, 1988; *Three Rival Versions of Moral Enquiry*. Duckworth, London, 1990; Charles TAYLOR, *Sources of the Self: The Making of the Modern Identity*. Cambridge University Press, Cambridge, 1989.

1. *Utilitarian theories*

Utilitarianism has, for a long time, been a very influential tenet in normative ethics. The attractiveness of the utilitarian approach derives from various sources: unlike deontological theories, it pays specific attention to the promotion of good. This good, unlike in ethical egoism, concerns other people, not only the moral subject. In utilitarian theories, the aim of morality is the maximization of utility, which is interpreted as the greatest possible balance of good over evil.[1] Furthermore, these theories usually define utility in terms of happiness or well-being, which connects utility with human desires and their fulfilment. The utilitarian good is a non-moral good, and things that can be placed in the categories of good and bad are taken to be mutually measurable and comparable in proportional relation to each other. Utilitarianism is a form of consequentialism: its aim is to establish states of affairs which, according to different criteria, represent what people prefer. The moral aspect enters the scene through intentional human activity: actions which further what people prefer are morally good, and actions which prevent what is desired are morally bad. Utilitarian models of moral reasoning are conceived as measuring procedures of a more or less technical nature. The morally best outcome is calculable from the premises given in the situation with the help of the formal procedure provided by the theory.[2]

The non-moral definition of the good and the element of mutual comparability have often been criticized by the opponents of utilitarianism. In the utilitarian view, there is nothing good or bad as such, but the goodness and badness of things is determined by the effect they have on (general) utility. Consequently, actions, deeds and policies that promote

---

[1] Frankena 1973, 34.

[2] Frankena 1973, 34–35, Williams 1987, 75–76.

this utility better than any alternative actions, deeds or policies in a given situation are morally good. This feature of utilitarianism has given its critics reason to claim that utilitarian theories are incapable of prohibiting actions, which in the light of other theories, are immoral. Another source of criticism is that utilitarianism ignores the fact that there are many different kinds of both non-moral and moral values, which cannot necessarily be evaluated with any one single measure, as the utilitarian models presuppose.[3]

In this chapter I will examine the theories of *Richard Brandt*, *Richard M. Hare* and *Derek Parfit*, who all claim to continue the utilitarian tradition. Brandt's theory is an attempt to show that morality can be connected with rationality by basing it on human psychology. Morality is rational, and rationally justifiable, as far as it can be connected with facts gathered about the human psyche. Hare for his part, continues the tradition of analytical ethics, contending that the conceptual study of moral concepts is the only solid basis for understanding morality correctly. In his normative moral theory Hare joins the utilitarian approach by maintaining that the aim of morality is the maximization of desire satisfaction. Derek Parfit holds that we actually cherish a false view of morality, which is connected with our way of understanding ourselves. The "true view" can be achieved through a revised understanding of our nature as persons. Adopting this revision provides both a justificatory basis for a utilitarian ethics and a motivation for being moral.

At first glance, these three theorists seem to have very little in common, yet they all claim to join the utilitarian tradition: for each, the purpose of morality is the maximization of (some kind of) desire satisfaction. The aim of this study is to analyze different moral theories through the concept of a person; and the question is whether we can find a joint conception of a person behind these utilitarian theories, or whether they all present versions of their own. If these theories are representative, we can ask whether there is a utilitarian conception of a moral person. Does the "utilitarian person" demonstrate something

[3] Williams 1987, 86–87,

more general about utilitarian moral theories? I will try to find an answer to these questions by studying each theory both as a theoretical model and as a normative device.

The present section will start with an examination of Brandt's theory, moving then to Hare, and completing the analysis with Parfit's theory. In the course of the scrutiny, attention will first be paid to the general structure of each theory, and an analysis of respective concepts of "person" will be based on this. The approach then allows for a comparison of the different features of the concept in different contexts and a more general analysis of utilitarian moral theories.

1.1. The "fully rational person" of Richard Brandt's theory

Richard Brandt is a meta-ethical *reductionist*: the task of moral philosophy is to translate moral language into a non-moral mode of speech. For Brandt this means that traditional moral issues are rephrased as questions concerning the rational.[4] Brandt seems to think that this reductionist approach is not merely a meta-ethical tool for moral theorists, but that it also has practical moral importance: it enables us to distinguish those cultural features which simply derive from traditional ways of thought from beliefs which can be based on a rational grounding.[5] Brandt contrasts his approach with intuitionism, which has long been an influential strand in moral philosophy. Intuitionism is an indefensible ethical position for it cannot offer any rational justification for moral views, thus making the basis of morality arbitrary.[6]

Brandt's redefinition of morality contains three stages: first, he shows

[4] BRANDT 1979, 2–3, 10, 14.

[5] BRANDT 1979, 1, 21–22.

how classical moral concepts and moral problems can be converted into questions regarding rationality; second, he presents a version of the theory of action which serves as a model for estimating the rationality of any action, and he finally transforms this model to cover the moral aspect and the criteria of moral acceptability. It is noteworthy that Brandt's theory is strongly psychological in orientation;[7] it explicitly attempts to establish a link between psychological facts and ethical norms, and can for this reason be classified as a *naturalist* theory.[8] Brandt connects moral questions on the one hand with questions concerning rationality, yet on the other hand he sees them as arising from human psychology. This makes Brandt a reductionist naturalist who attempts to make the rational understandable through human psychology. Successful

---

[6] Brandt distinguishes two types of appeal to intuitions: the *method of appeal to linguistic intuition*s and *ethical intuitionism*. The first type of intuitionism assumes that "there is some expression synonymous with any normative statement in its ordinary use, which differs from the original enough to make clear what has to be done to confirm or justify the original statement". Brandt 1979, 4. Here sentences containing moral terms are identified with their respective synonymous paraphrases by appeal to our linguistic intuitions. This step does not guarantee any reliable results because normative words are so vague in their meaning that linguistic intuitions cannot grasp anything firm, or absolute. A further reason for abandoning this approach is that intuitions are not a reliable basis for guidance in normative reflection: language might embody confusing distinctions which we fail to notice if we just refer to our intuitions. Finally, linguistic intuitions cannot provide an auxiliar for discerning many problems which we, on reflection, find essential, and they do not suffice for raising questions which we, on reflection, want to raise. Brandt 1979, 4–10. The form of ethical intuitionism Brandt opposes maintains that having intuitions simply means that "we do make normative statements sincerely, and believe them in some ordinary sense". Brandt 1979, 17. Ethical intuitionism goes astray because it cannot provide us with any grounds or justification for the reliance on intuitions. In addition, the fact that our intuitions are usually an amalgam of what we have learned and adopted from various sources during our lives should make us suspicious of them as a basis for rational action. Intuitions also vary from society to society, and from culture to culture; even the intuitions of individuals change depending on their moods or situation in life. Instead of relying on intuitions, ethics should "step outside our own tradition somehow, see it from the outside, and evaluate it". Brandt 1979, 21. Appeals to intuitions fail to provide us with such a neutral point of view. Brandt 1979, 16–22.

[7] The presentation of the psychological theory needed for defining the morally acceptable accords for one third of the 335-page book.

[8] The redefinition of moral questions is accomplished by means of conceptual analysis, and the aim of this reformulation seems to be to subject the traditional moral questions to scientific or observational inquiry. Brandt 1979, 2–3.

moral philosophy finally cancels itself out, since moral considerations are brought back to the rational, which is, again, fully explicable through human psychology.

1.1.1. THE STARTING-POINT OF MORAL PHILOSOPHY

Perceived facts, logic and scientific knowledge form the basis of ethical, as of any other inquiry.[9] Applied to moral philosophy, this involves finding something common to all people regardless of their particular frameworks, cultures and social situations.[10] Such a basis is a necessary condition for any neutral, unbiased approach to ethical questions.[11] A ground of this type can be found, Brandt maintains, since all people act and make decisions, all people have wants, desires and aversions irrespective of their social, cultural, and historical backgrounds. Furthermore, every society has some moral system for regulating the behaviour of its members. All people find it important to answer certain questions, and especially the question as to how far actions, desires, and moral systems can be criticized by appealing to facts and observation.[12] This is the ground that unites all people irrespective of time and place, and this is

---

[9] BRANDT 1979, 2, 16, 21–22.

[10] BRANDT 1979, 10. Finding a common basis does not mean that we have to show that actual moral systems share some mutual, universal conception of moral reasoning. In his former work (*Ethical Theory.* Prentice Hall, Englewood Cliffs, New Jersey, 1959.) Brandt maintains that all people actually have and apply a similar method in their moral reasoning. Again, this method, if elaborated, can be developed into a universally valid ethical method. Brandt dismisses this view as mistaken; we cannot say that people actually exercise a similar method in their moral reasoning. See BRANDT 1979, v.

[11] BRANDT 1979, 10, 185.

[12] "Every person acts, or makes decisions. Every person has desires, aversions, aspirations. Every society has some kind of moral system. So there are some topics about which everyone will want a question answered; and one question everyone will want answered is how far actions, desires, and moral systems can be criticized by appeal to facts and observation." BRANDT 1979, 10.

the basis for Brandt's moral theory.

What then does Brandt's list of basic human characteristics express? Brandt calls his selection of human features *facts*. He does not, however, arrive at them by any empirical analysis, which raises the epistemological question of what sort of facts they actually are. They are certainly not facts of the same kind as "all human beings have lungs and a heart", or "all human beings are mammals", but something else. It seems that Brandt's starting point — "all people have desires", "all people act", "all people ask how far actions, desires, and moral systems can be criticized by appealing to facts and observation" and "every society has some kind of moral system" — describes a distinction between the *morally* relevant and other human considerations. We can discern two types of "facts" in Brandt's list: those that refer to human intentionality in general and those that belong particularly to the moral realm. As part of the intentional, the starting point of morality is a *desiring acting agent*. This desiring acting agent moves to the moral realm when she ask "how far actions, desires, and moral systems can be criticized by appealing to facts and observation" and tries to give an expression to this in the form of a societal moral system. In this way, the facts Brandt uses as his starting point present an initial description of the moral as part of the intentional and gives an outline of what distinguishes morality from other kinds of intentionality. It is noteworthy that the moral point of view also gives us an initial definition of a *moral person*, because according to this theory the morally relevant is given through what makes a person enter the moral realm.

According to Brandt, everyone wants to find an answer to the above question. This involves actually asking for the *rationality* of one's actions, desires and moral systems; and what all people in fact want to know is whether their actions, desires and moral codes are rational.[13] But this is, Brandt maintains, not different from asking traditional moral questions, such as "what is desirable in itself?", or "what is the best thing to do?".

---

[13] Brandt employs the term 'rational' to "refer to actions, desires, or moral systems which survive maximal criticism and correction by facts and logic". The initial starting point thus acquires a new form, namely: "How far can actions, desires, and moral systems be said to be rational?" Brandt 1979, 10.

Consequently, instead of inquiring into the morality of actions one can simply examine their rationality.[14]

To legitimize this reduction, Brandt argues that people who ask what is the best thing to do actually want to know which action is more worthy of being chosen than any other. Moreover, identifying this action means that people must know what they would choose if they had scrutinized the causal factors responsible for their choice in all possible ways. In this manner, the estimation of feasible actions is reduced to an inquiry about their rationality. Brandt defends this redefinition of the "morally good and right" in terms of the rational by saying that defining the rational thing to do covers all important issues raised by the more traditional questions concerning the goodness and worth of action, and additionally, it even encompasses a number of issues that are left out with traditional questions.[15] Furthermore, "rational" is a concept with a recommendatory function: people tend to act in a rational manner, and when they are shown that something is rational rather than something else they are inclined to follow this advice.[16]

Brandt's redefinition programme, if it succeeds, *reduces morality to rationality*, and translates moral language into questions concerning rationality without a residue. Brandt seems to think that the concept of rationality is not as arbitrary as that of morality, but that it is easier to agree on what is rational than on what should be one's moral basis for action. If moral dilemmas can be reduced to questions of rationality this makes it possible to apply criteria of rationality to moral issues and thus enables a better solution to ethical disputes to be found.

There is still another side to Brandt's redefinitory programme: desires

---

[14] BRANDT 1979, 14.

[15] BRANDT 1979, 14–15.

[16] Brandt regards this move as unproblematic, there are even several reasons in favour of substituting traditional moral concepts, such as "good" and "right", with "rational". First, the meaning of traditional notions is controversial, and indefinite. An exact definition of "rational" will clear up this problem. The second reason for adopting the reformulation is that "rational" is not only a good descriptive concept, it also performs a recommendatory function: people want, and they understand that it is always favourable for them, to act rationally. BRANDT 1979, 15.

and actions determine the realm of intentionality in which morality is defined. As we have seen, this definition includes the basic features of the Brandtian moral person. Now, we can make a further distinction between a *natural person* and a *moral person*. The difference between these two is that the desires and actions of a natural person arise spontaneously as they emerge from the psychological and cognitive mechanisms of the person. As such, these actions are intentional, and thus potentially moral, but they are not morally qualified. The moral person is different. Like the natural person, she is an intentional agent, but her desires and actions have been qualified by criticism which causes them to fulfil the criteria of rationality. She is, thus, a rational person. For a further analysis of Brandt's theory from this perspective we need to examine closely the concept of rationality; I will now study Brandt's method for estimating the rationality of actions and desires.

1.1.2. The method of cognitive psychotherapy

What is action? Brandt's theory of action can best be described as a version of the traditional, teleological theory of action, which takes the form of a *practical syllogism*: from the two premises, *A wants that p*; and *A believes, that not p if not q*; the consequence, *A starts doing q*, follows. In other words, a person starts performing an action *q* to attain a desired goal *p*.[17] According to the traditional view, evaluating the rationality of an action

[17] Brandt refers to his theory of action as the 'cognitive theory of action', and he wants to give it a behavioural basis grounded on modern psychological research. For the full exposition of the theory, see Brandt 1979, 46–59. The following quotation reveals the core of Brandt's theory: "[...] action-tendencies are a multiplicative function of valences (occurrent desires and aversions), and hence [...] an action-tendency is always zero in magnitude if there is no valence attached to the contemplated action itself or its expected outcome. [...] If some philosophers have thought, as some seem to have done, that a person can do his duty even if so doing is not positively valenced for him (or failure negatively valenced), perhaps 'out of respect' for duty in some sense, they were wrong; and their psychology of morality needs basic revision." Brandt 1979, 66–67.

only concerns the second premise of the syllogism, that is, the rationality of the means, not the rationality of the ends. The question is, then, whether an agent is rationally justified to perform an action *q* inthe face of her beliefs concerning the connection between the desired goal *p* and the action *q*. Consequently, an action is rational if an agent is rationally warranted in believing *q* to be a proper means for attaining *p*. Now, Brandt maintains that his theory expands the claim for rationality further to include the rational justifiability of the first premise, i.e., the rationality of the ends. There are, namely, criteria for evaluating the rationality of a *want* of an agent and the rationality of a *desire* upon which particular wants are based. This evaluation, Brandt believes, makes it possible to estimate the rationality of both means and ends by subjecting both premises of a practical syllogism to "maximal criticism based upon facts and logic". Premises which have undergone this criticism, and which have been modified to accord with the requirements of logic and facts, are *rational*. This procedure makes the two premises of practical syllogism rational, and for this reason, the conclusion deduced from them must of necessity also be rational: an action, based on rational wants and desires and achieved through rational means is rational.[18]

Brandt employs a *cognitive theory of action* to explain how the rationality of desires and actions can be estimated. The evalution has two phases: the *first approximation of rationality* only concerns the (relatively unproblematic) notion of the rationality of chosen means for a given end. This

[18] "It is convenient to distinguish three senses of 'rational', for use in different contexts. First, I shall call a person's action 'rational' in the sense of being rational to a first approximation, if and only if it is what he would have done if all the mechanisms determining action except for his desires and aversions (which are taken as they are) — that is the *cognitive inputs* influencing decision/action — had been optimal as far as possible. [...] Second, I shall call a desire or aversion 'rational' if and only if it is what it would have been had the person undergone *cognitive psychotherapy* [...] Finally, I shall say that an action is 'rational' in the sense of fully rational if and only if the desires and aversions which are involved in the action are rational, and if the condition is met for rationality to a first approximation." BRANDT 1979, 11. I have called Brandt's 'rationality to a first approximation' *rationality concerning the means*, and the rationality determined by cognitive psychotherapy *rationality concerning the ends*. What Brandt calls 'fully rational' is, according to the terminology of practical syllogism, the conclusion of practical syllogism when rational means are applied to attaining a rational end.

stage concentrates on examining the rationality of the beliefs upon which an agent bases her action. The *second approximation of rationality* explicates grounds upon which wants and desires can be criticized, and hence, when one is warranted to speak of rational ends.[19] The first and the second phase together form a procedure called *cognitive psychotherapy*.[20] Actions which square with the requirements of cognitive psychotherapy are called *fully rational*.[21]

To understand Brandt's cognitive theory of action, we must return to his definition of the moral as that which is rational. The first approximation of rationality — rationality of chosen means for a given end — coheres with the standard notion of rationality, whereas the second approximation of rationality — or rationality of ends — includes the moral realm. An agent is, then, *fully rational* if her action is both rationally and morally flawless. Brandt joins the tradition of moral philosophy which regards desire as the driving force of human action. Unlike many of the proponents of this view he does not, however, share the opinion that desire is blind to reason.[22] On the contrary, Brandt's central idea is that reason can not only determine the best means for our desired ends, but also qualify our natural desires to meet the demands of rationality. This is actually just what moral theory helps to achieve.

The psychological framework defining which wants, actions, and

---

[19] "[...] the question 'What is the best thing to do?' can helpfully be reformulated as 'What would I do now if my decision processes had been subject to criticism by facts and reason to a maximal extent?' or, given that terms are suitably defined, as 'What is the rational thing for me to do?'. [...] decisions are a function of both cognitive factors and desires/aversions, and that therefore an answer to the question what is the fully rational thing to do must come in two stages. In the first stage, we concentrate on determining what is the rational thing to do, as a first approximation, concentrating on the cognitive factors, and taking for granted that the desires/aversions are given, not needing criticism. In the second stage, the desires and aversions themselves are criticized, and thereafter it may be determined what fully rational action would be." BRANDT 1979, 46. See also BRANDT 1979, 149.

[20] BRANDT 1979, 113.

[21] BRANDT 1979, 46, 88.

[22] The thought that desire is blind to reason and reason is inapt to act is usually connected with David Hume; see his *Treatise of Human Nature*, Book II, Part III, Section 3, "Of the influencing motives of the will".

desires are rational plays a dual role in the theory. First, it describes what is involved in human desiring and acting. Second, it determines the criteria for their rationality, i.e., of the morally good. In this way, the psychological theory connects facts and norms within the moral theory.[23] On the one hand, the rationality of desires and actions is based on this psychological framework. Here the psychological theory has a *descriptive function*, while it gives an account of what kinds of desires and actions can be characterized as rational on the basis of the facts of the human psyche. On the other hand, the psychological theory also exercises a *normative function* since it provides the criteria which people must use for determining which of their desires and actions are rational, and hence, morally good. To explicate the conception of "fully rational" further, we need more information of how the rationally and the morally correct are defined. For this purpose we must examine Brandt's theory of wants.

Wants and *aversions* are central concepts in Brandt's cognitive theory of action. Roughly, the things that a person wants to attain have a positive valence, and things she wants to avoid have a negative valence for her.[24] Human action is intentional in the sense that an agent cannot be said to act if what she does is not done for the purpose of realizing some wanted, positively valenced state of affairs. An agent can fail to reach a wanted goal for any number of reasons, but she has to believe (however nominally) that the action taken will forward her goal.[25] There are two kinds of mistakes in reasoning which weaken the rationality of an action; first, errors which concern the gathering, or use of information relevant

---

[23] Brandt explicitly maintains this to be his aim, and: "[...] even if these generalizations turn out to require modification, at least a demonstration of how psychological theory can be used to establish normative principles will be worthwhile in showing how someone can turn, to the same use, the theories of the future. Of course, it is not news that philosophers should think they can squeeze normative conclusions out of psychology: they have been trying to do this since Epicurus, and Hume is a prime example of the psychologizing of moral philosophy." BRANDT 1979, 2.

[24] "People *want* events or situations of a certain sort to obtain at some time or times; or they are *aversive* to events or situations in the sense that they want them not to obtain at some time or times. We can speak, meaning the same, of an event or situation obtaining at a certain time as being positively (negatively) *valenced* for a given individual at a certain time." BRANDT 1979, 25.

for the realization of the action, and second, mistakes that derive from giving one's present preferences, and the consequences of action in the near future unwarranted attention in comparison to the ones in the less immediate future.[26] It is a *necessary condition* for the rationality of action that one does not commit these errors; or, in other words, an action is never rational if the best possible means, facts and logic considered, are not used for attaining one's given end.

In this connection, Brandt speaks about rationality of actions but, to be precise, he should speak about rationality of beliefs concerning actions as means for a given end. The inaccuracy of his terminology indicates a deeper ambiguity in the theory. The rationality of means seems to indicate that an agent, in order to be justifiably called rational, must necessarily base her action on beliefs which correspond with the state (or states) of affairs in the outside world, because otherwise it is improbable that she would achieve her goals. Interestingly however, Brandt defines the criteria of rationality, not by linking them with states of affairs, but by making them dependent on the psychological conditions of an agent, which means that he establishes *psychological criteria* for rationality.

Even if an action is flawless according to the rationality of means this

[25] "What is to count as an 'action'? It is convenient (but not necessary) to limit 'action' to intentional bodily movements and mental occurrences: like waving one's arm, putting on one's hat and so on. [...] For the most part, however, deliberation is concerned with the selection and execution of a sequence of acts comprising a complex plan. First there is the selection of a sequence (its later stages left a bit imprecise); then the initiation of its first stage; and finally the monitoring, maintenance, or modification of its later stages. Obviously the selection and execution of such a plan involves many different bodily movements, connected by the belief that the sequence of actions will accomplish certain results." Brandt 1979, 47–48. See also Brandt 1979, 66.

[26] First, an action cannot be rational if it is based on an incomplete set of possible options. An agent notices the alternative courses of action only if she has all the relevant information about the situation and uses her information in an adequate way. Brandt distinguishes two degrees of rationality at this point: an action is *subjectively rational* if it is based on the full use of beliefs rationally supported by the evidence the agent actually has at the time. A fully rational agent ought to aim at objective rationality, though; in this case the agent utilizes all relevant information available at the time of acting, including such pieces of information not easily available. The second test of rationality examines how well an agent takes different, possible outcomes of different actions into consideration. Such mistakes may derive from insufficient information, or from neglect of information, or from a misapprehension of it. Brandt 1979, 70–87.

does not make it rational, for it is only a necessary condition of rationality. In addition to the rationality of means, one must determine the criteria forthe rationality of an agent's wants and aversions, because this alone can establish the *sufficient conditions* for rationality of action. The psychophysical constitution of human beings means that some things give a pleasant feeling, while others cause uncomfortableness.[27] In particular, the satisfaction and deprivation of basic human needs have this effect: it is pleasant, let us say, to be nurtured and cared for. The pleasure caused by the satisfaction of basic human needs is innate, and the pleasantness of all other experiences of pleasure derives from this basis.[28] Most experiences are neutral, but they acquire a pleasant or unpleasant colouring as they are associated, through the mechanisms of conditioning, with innately pleasant or unpleasant experiences.[29] When an experience *B* has become pleasant for someone by having repeatedly occurred in association with some already pleasant experience *A*, the experience *B* then feels pleasant independently of *A*.[30] The valence of an experience attached to some object of desire or of aversion does not change over time; instead, it can only be altered through counter-conditioning. Thus, a positively valenced experience will turn unpleasant if it constantly occurs in the context of another, unpleasant experience. The mechanisms of conditioning and counter-conditioning enable people to influence their own behaviour by moulding the desires and aversions upon which behaviour is based.[31]

---

[27] Brandt defines the experience *E* of the person *P* to be 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." BRANDT 1979, 40–41.

[28] BRANDT 1979, 90.

[29] BRANDT 1979, 90–92.

[30] BRANDT 1979, 94–95.

[31] BRANDT 1979, 103–105.

Should a person change her desires and aversions? What direction ought this change to take? Brandt maintains that there are certain desires which are rational and that all rational people find the same things pleasant to a certain degree. Similarly, there are desires (and aversions) which are irrational, or irrational at least when they exceed a certain limit. A person should not, and cannot, change her basic innate desires and aversions. They belong to the human constitution as products of evolution and they have, as such, survival value.[32]

Brandt's definition of the rational implies that all non-extinguishable desires and aversions are rational; they cannot be changed no matter how much we allow facts and logic to influence them.[33] We should, however, change those of our desires and aversions which do not meet the requirements of rationality, that is to say, which change by means of facts and logic.[34] How can we then identify which of our desires are irrational?

Brandt presents two distinct procedures — one positive, the other negative — for evaluating the rationality of desires. The positive approach, called *cognitive psychotherapy*, is a method for moulding and developing desires by means of facts and logic.[35] The negative procedure offers a test for recognizing irrational desires: a desire is irrational if it involves one, or more, of the four mistakes listed by Brandt.[36] These two

---

[32] Brandt 1979, 113. See also Brandt 1988, 72–73.

[33] "If a desire will not extinguish, then it is not irrational. This result is consistent with the general view that a desire (etc.) is rational if it has been influenced by facts and logic as much as possible. Unextinguishable desires meet this condition." Brandt 1979, 113.

[34] An irrational desire causes inconsistency and dissonance among a person's total set of desires which is analogical to the inconsistency and dissonance deriving from a false piece of information in a person's cognitive system. Incongruity within a system of desires is unpleasant to the extent that it can inhibit action. All irrational desires have received so high a valence that the satisfaction of these desires incapacitates a person from satisfying her other desires. It is, however, in all people's interest, no matter who they are and what their aim is in life, that their sets of desires does not include any such irrational elements. Hence, if someone finds her system of desires inconsonant, she naturally tries to modify it, as she will modify her beliefs when she notices that her cognitive system includes mutually contradicting, or inconsistent components. Brandt 1979, 156–159.

[35] Brandt 1979, 113.

methods provide the necessary and sufficient conditions for the rationality of desires: the negative test determines the *necessary condition* and the positive approach defines the *sufficient conditions*.

In cognitive psychotherapy an agent examines her prevailing desires in the light of acquirable, relevant information by repeating it in her mind as vividly as possible when she makes a decision.[37] Cognitive psychotherapy creates, strengthens and develops rational desires and weakens, or even extinguishes irrational desires. Vividly and repeatedly presented information, relevant for the agent, moulds her desires, and hence affects her action.[38] Consequently, a desire or an aversion is rational by definition if it has passed the test of cognitive psychotherapy, or if cognitive psychotherapy has created it. Accordingly, if cognitive psychotherapy has no effect on a desire it is rational. Such desires fulfil the sufficient condition of rationality: they have been maximally exposed to criticism stemming from logic and facts.[39]

The preceding analysis provides the necessary material for a further analysis of the link between the rational and the moral.[40] All people

---

[36] Mistakes, which deplore the rationality of ends and cause irrational desires, centre around false or mistakenly used information about the world. To take an example, a deprivation a person has suffered during her childhood can make her attach an incorrectly great valence to an object she connects with the deprived desire. Irrational desires and aversions also stem from the cultural heritage. BRANDT 1979, 115–126.

[37] BRANDT 1979, 113. By the acquirable, relevant information, Brandt means scientific and other information the agent can acquire, plus both inductive and deductive reasoning. BRANDT 1979, 111–112.

[38] See, page 27.

[39] "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. [...] One implication of our definitions may be surprising. It arises from the fact that some valences, or dispositions to enjoy something, may resist extinction by inhibition and anything else, since they have been so firmly learned at an early age. By my definition these qualify as rational. For I use 'rational' as the contradictory of 'irrational' and have defined an 'irrational' desire (etc.) as one that would extinguish after cognitive psychotherapy. If a desire will not extinguish, then it is not irrational. This result is consistent with the general view that a desire (etc.) is rational if it has been influenced by facts and logic as much as possible. Unextinguishable desires meet this condition." BRANDT 1979, 113.

belong to some society, or group, and they usually adopt the moral code of their society. Existing moral codes are not, however, necessarily such that their commensurate desires and wants automatically fulfil the criteria Brandt's theory sets out for rational wants and desires. Any person who wants to be rational has to evaluate the principles and rules she has learned and adopted, and the desires and wants which accord with them, in the light of cognitive psychotherapy.[41] In an actual situation the decision of an agent concerns not only the action she performs, but she also decides, indirectly, which moral principle she will follow. Her decision is rational, i.e., morally justified if both the goal of her action stated in the principle, and the steps she uses for reaching her goal are rational.

This analysis shows how the psychologically defined rational determines the ethically acceptable: the estimation of moral worth concentrates on the *psychological qualities* of an agent, not on the characteristics of her deeds. Brandt's theory reduces morality to a quasi-natural theory of cognitive psychotherapy which defines morally right actions as actions which a *fully rational person* would perform, that is to say, an agent whose desires, wants, and beliefs have survived maximal criticism stemming from facts and logic.[42]The psychological theory Brandt uses for describing the emergence of human desires and aversions and their effect on

---

[40] See, page 27.

[41] One of Brandt's central aims is to develop a method of moral reasoning which enables people to transcend their own tradition. BRANDT 1979, 21–22.

[42] BRANDT 1979, 126–127. Brandt admits that the method of cognitive psychotherapy allows a variety of outcomes which must all be rendered the status 'rational', but which are, however, incompatible with each other. It is also possible that a person using cognitive psychotherapy at some moment arrives at a conclusion different from the resolution she would find rational at another moment of time. In Brandt's view, this does not weaken the theory; even rational desires can change with time. BRANDT 1979, 181–182, 188, 200–203. There is also another difficulty, which can complicate moral evaluation. There are, namely, no criteria for deciding which of the two would be better: to foster few very strong desires and aim at their satisfaction, or to keep and minister to a multiplicity of different desires, the satisfaction of which ensures a richer life for the agent. If an agent gains boundless pleasure from satisfying one single desire, and forsakes everything else in life because of it, is she being irrational? Brandt does not explicate any criteria for solving this problem, but there are, however, hints which suggest that he would not consider it to be rational for a life to be structured around the satisfaction of one single end.

human action offers a criterion for the rationality of desires — aversions as well as actions. Consequently, if someone's psychological setup is flawless, she necessarily chooses a rational course of action, which is by definition also a morally right course of action. Brandt's solution is problematic, because there are no material or substantive criteria of moral correctness outside an agent's psyche. If an agent has subjected her desires and the actions based on them to cognitive psychotherapy, both are supposedly morally irreprehensible, no matter what they are. This problem is, however, not unique to Brandt's theory; but it features in all theories in which moral correctness depends solely on the defined qualities of an agent.[43]

This is the outline of Brandt's theory and we can now examine the concept of a person which is embedded in it. As we have seen, Brandt's definition of the morally relevant already includes a definition of a moral person as an intentional agent who asks for the rationality of her actions and desires to be taken into account.[44] We have also noticed that we can make a distinction between a *natural* and a *moral person* in the context of the theory.[45] Of these two the second has undergone cognitive psychotherapy and her desires and actions are influenced by that, whereas the first has not submitted herself to such treatment. As has already been

[43] The so-called *agent-relative* theories concentrate on defining moral goodness in terms of the qualities of the moral agent. This makes morality less dependent, or completely independent of moral rules which are regarded as insensitive to the complexity of the moral situations in real life. These theories usually give a central role to such concepts as 'virtue' and 'character'. The agent-relative nature of these theories becomes explicit in situations when a moral quality is evaluated differently depending on whose quality it is, what part it has in a person's character, or life, etc. These theories also evaluate actions in an agent-relative way, the quality of an action depending on the quality of the agent. In their most extreme forms these theories claim that a morally bad person cannot perform morally valuable actions while her character is defective, and is as such unable to produce anything morally good. The same applies, conversely, to a person with a good character, she cannot actually perform anything that is morally bad. To avoid this difficulty a moral theory cannot be completely agent-relative, but there have to be criteria in the theory for at least distinguishing the deeds that are always morally prohibited irrespective of the agent. This is just how the problem is avoided in the traditional virtue theories; see, e.g., EN III, 1, 1110a.

[44] See, page 26.

[45] See, page 27.

pointed out, a moral person is necessarily rational. Brandt's description of the morally relevant defines morality as a subclass of intentional actions, which follow the form of the practical syllogism. It is noteworthy that in Brandt's formulations the moral aspect enters the realm of intentionality through the agent, as she frames her desires and actions subject to the criticism of rationality. Apart from this, there is nothing in Brandt's formal definition that distinguishes morality from other kinds of intentionality. For the practical purposes of moral reasoning Brandt completes his model with a normative device. This role is given to cognitive psychotherapy, which qualifies natural desires, making them morally acceptable. Because morality can be reduced to rationality, cognitive psychotherapy first makes a person's desires and then her actions rational.

In Brandt's theory, moral actions and desires form a subclass among rational actions and desires; it is thus rational to be moral. This formulation has a curious consequence: if we want to be rational agents, and Brandt thinks that we all inevitably do, we will have to consider all our desires and actions in the light of cognitive psychotherapy, because it determines whether a planned action falls into the category of the rational or not. But does this not, instead of reducing morality into rationality, make all intentionality morally relevant? If we adopt Brandt's redefinition of the moral terms as the basis of our moral thinking, the moral point of view "takes over" all other aspects of human life through the fact that we must apply cognitive psychotherapy to every want and desire, and to every planned action. All intentional action becomes moral. The moral person completely replaces the natural person. Brandt's theory, if taken seriously, represents a form of "ethical imperialism" which demolishes the non-moral perspective on human life; every issue becomes a moral question.

1.1.3. THE MORAL CODE OF A FULLY RATIONAL PERSON

The method of cognitive psychotherapy determines which desires, wants and actions are morally correct. Estimating singular actions by cognitive psychotherapy is evaluation of the act-utilitarian type. Brandt is, however, not an act-utilitarian but rather an exponent of rule-utilitarianism: moral philosophy should offer grounds for accepting a moral code. What is needed, then, is a link between the moral justification of a singular action, provided by cognitive psychotherapy, and a moral code. Brandt constructs the connection through the concept of a *fully rational person*. Consequently, the task of moral philosophy is to find an answer to the question *which social moral code, if any, would a fully rational person tend to support*. By formulating the central moral question in this manner, Brandt demarcates his theory from three other solutions: first, moral theory must search for the *social moral code*, not concentrate on defining the features of the morally right actions; second, *tend to support* is a more intelligible concept than *choosing* a moral code, a notion much used especially in social contract theories;[46] and third, it is better to speak about *a fully rational person* than about an ideal observer.[47]

Brandt maintains that a moral code is not solely nor even primarily, an intellectual matter, but that its emotional elements are central.[48] To sup-

[46] Brandt criticizes theories, and especially John Rawls' theory, which employs the concept of 'choosing a moral code'. Brandt regards these formulations as problematic: to speak of a moral code as an object of one's choice presupposes that the choice can cover, among other things, the strength of the support given to the moral code, and the mechanisms of socialization that transfer a moral code from generation to generation. In order to avoid these difficulties, Brandt prefers the phrase 'tend to support a moral code'. It is not certain, however, that Brandt's 'tendency to support a moral code' actually means anything else than just choosing a moral code. Brandt's criticism of the Rawlsian 'choice' focuses on the point that Rawls seems to ignore certain psychological facts which limit the variety of moral codes open for choice. 'The tendency to support' amounts to restricting the set of possible moral codes to those not involving a contradiction with what Brandt supposes to be psychologically viable for human beings to fulfil. If these restrictions are taken into account, one can speak of 'choosing' a moral code without inflicting violations upon Brandt's system. Thus, a moral code must not impose costs and difficulties impossible or hard to bear. It must not set demands that are more costly than the benefits that the moral code brings with it are worth. Further, it must be adaptable, learnable, and socializable. BRANDT 1979, 188–190.

port a moral code involves a person having the intrinsic motivation to act in certain ways, and that she feels guilt when she does not succeed in acting according to this motivation. Further, a person holds that action types that accord with her intrinsic motivation are important and justified, and is able to articulate these basic motivations.[49] This formulation is problematic. Brandt's definition is formal; now, if someone "supports a moral code" which inhibits, for example, killing the innocent, does this imply that this person simply "has a motivation for not killing an innocent", or can we draw a stronger conclusion, namely, that she tries to direct her behaviour according to the rule forbidding such killing? What does "supporting a moral code" actually mean? On the one hand, Brandt speaks about intrinsic motivations, on the other he seems to think that moral codes consist of articulated moral rules. Does being a morally good person involve following certain norms, or just having certain motivations? Brandt offers no explicit answer to this question, he simply seems to presuppose that being motivated in a certain way, supporting a moral code, and refraining from certain deeds are all one and the same thing. This ambiguity indicates, however, that the move from a set of fully rational actions to a developed moral code is more problematic than Brandt acknowledges. The disposition to feel guilt, central to Brandt's definition of a moral code, is itself not a desire or a want.[50] How

---

[47] Several utilitarian theories use the concept of an *ideal observer* to guarantee non-partiality between people. Brandt does not accept the notion, for knowing what an "omniscient, omnipercipient, disinterested, dispassionate, but otherwise normal person" would do does not help *us* in our moral dilemmas. Brandt 1979, 22. A *fully rational person* is a better concept; although a fully rational person makes use of all available information, she is not omniscient or onmipercipient. She does not have to be disinterested or benevolent as long as her wants, desires and beliefs have been qualified by cognitive psychotherapy. This requirement has the same effect as the ideal observer theory and, additionally, it secures the rationality of desires which direct an agent's actions. Moreover, a fully rational person is much more "normal" than an ideal observer. Brandt 1979, 225–228.

[48] Brandt 1979, 170–171.

[49] Brandt 1979, 169–170.

[50] "People do not like to have others disapprove of their behaviour; and, if a person knows others disapprove of what he did, he will feel some discomfort which we could also call 'feeling guilty'." Brandt 1979, 167.

do such moral dispositions fit cognitive psychotherapy? If they form an innate part of the human psyche, should Brandt not include them in his psychological framework? Or is the tendency to feel guilt a distinctively *moral* feature?[51] If it is, then Brandt's reductionist project fails; it does not allow the reduction of the moral to the rational.

In Brandt's view a social moral code is a more important concept than an individual's moral code. Accordingly, a social moral code is directly derivable from individual moral codes represented in a society, while in a given society individuals' moral codes will largely coincide.[52] Constructing a social moral code means listing first all types of behaviour enjoined or prohibited by the moral code of at least one person. One then picks up the features included in most people's moral codes, and calls this the social moral code of a given society.[53] This definition of a social moral code seems odd against the background of Brandt's original question "which social moral code, if any, were a fully rational person tend to support". If the social moral code is constructed in this manner, what does supporting the code mean? Does Brandt simply maintain that a fully rational person *wishes* that the majority of her society has adopted a certain type of an individual moral code, or that they are motivated in a certain way? The definition for social moral code is descriptive, simply recounting the most typical attitudes or opinions concerning moral behaviour in a society; but to "support a moral code" also requires that the code has a regulative and prescriptive role in the person's thinking. The ambiguity in Brandt's formulation makes it difficult to discern what is really at issue here.

What kind of a moral code, if any, would a fully rational person then

---

[51] What BRANDT (1979, 167) states about feelings of guilt suggests that there is a distinctly *moral* feature in the theory: "Knowledge that others will disapprove increases the effectiveness of the moral code as a deterrent to behaviour; moreover, it plays a part in the acquisition of an individual's own moral code, through the process of conditioning. But it is confusing if we count a person's tendency to feel uncomfortable when he knows others disapprove as a part of *his* individual moral code — that is precisely what it is not. *His* moral code is evinced by his autonomous guilt-feelings — those arising from failure to act in accord with his own moral motivation."

[52] BRANDT 1979, 172.

[53] BRANDT 1979, 172–173.

support?[54] Brandt maintains that all fully rational persons would agree upon the fact that it is more rational to support some moral code regulating behaviour than not to support any such code. Likewise, all such persons would agree that some moral codes are more rational than others.[55] Despite the consensus on these matters, the moral codes fully rational persons support do diverge from each other. This stems from differing desires: even fully rational people attach different valences to different objects of desire.[56] Here we notice how desires, which are crucial for Brandt's definition of the morally relevant, also determine the content of a person's moral code.

The variable which determines the set of moral codes supported by fully rational moral agents is the degree of the agents' *benevolence*.[57] Here we face a similar problem to the one we encountered when we discussed the status of guilt in this theory: it is difficult to classify benevolence either as a want or a desire. How does benevolence relate to the desires which regulate a person's actions? Is benevolence caused by a desire to be benevolent? If so, should we not conclude that a person's desires do not determine her benevolence, but that her benevolence determines, at

[54] Why would anyone be interested in knowing which moral code she support sif she were a fully rational person? Brandt's response is that no one wants her moral code to be solely based on intuitions shaped by the tradition within which she lives, and to be coherent only in the sense that her intuitions are in unison with each other. Instead, each of us wishes her moral code to be "true", that it is, as far as possible, shaped by the requirements of facts and logic. To recognize a moral code that a fully rational person would support involves just this. The moral code a fully rational person would support does not rest upon factual errors, conceptual mistakes or misleading arguments, which serve as a recommendation in favour of supporting that moral code. Recognizing the moral code of a fully rational person justifies this moral code. BRANDT 1979, 185.

[55] BRANDT 1979, 200–203.

[56] BRANDT 1979, 200–201.

[57] Benevolence is a rational feeling. It is formed by a very early conditioning, therefore it is difficult, if not impossible, to extinguish it even by counter-conditioning. Brandt maintains, without giving a further reason for his conviction, that one should develop and increase one's degree of benevolence. If one does not become more benevolent by the process, one should not be morally held to blame: such people just have not developed the feeling when they were infants. BRANDT 1979, 139–140, 143–146. Brandt joins the British tradition of moral philosophy which has treated benevolence as a natural and innate feeling central for moral behaviour; see David HUME, *An Inquiry Concerning the Principles of Morals*, sec 9.

least partly, her desires? It seems that the concept of benevolence implicitly imports some concept of the *moral* good into the theory, something Brandt has definitely strived to avoid by his formal account of the good in terms of the psychologically defined rational.[58] This is a sign of the impossibility of Brandt's project; the naturalist reduction of morality does not succeed.[59]

To return to Brandt's theory, one might arrange different versions of a fully rational person's moral codes on a continuum, at the one end of which we place the moral code of a perfectly selfish agent, and at the other the code of a perfectly benevolent agent.[60] Both are theoretical options within whose boundaries all moral codes fulfilling the criteria can be placed. In reality, it is as unlikely to find a person wholly lacking benevolence, as it is to meet anyone perfectly benevolent. The two extremes of the rational moral codes also constitute another important division in Brandt's theory. The moral code of a fully selfish rational agent presents the category of the morally obligatory; it is a minimum set

---

[58] Brandt defines benevolence thus: "The term seems to encompass both motivational and affective elements; so I shall say a person is benevolent if it is a relatively permanent trait of his personality that (1) he is intrinsically motivated to produce happiness or welfare in others and to avoid decreasing it, (2) tends to be pleased if informed that someone has moved to a higher level, and (3) tends to be displeased if informed that someone is unhappy or not well off or has moved to a lower level. So defined, it is both a disposition to like or dislike something, and also a related desire or aversion." BRANDT 1979, 138. Brandt's formulation matches with the psychological theory he is advocating as the basis for moral theory, and benevolence is defined without a reference to moral characteristics. The moral aspect is, however, hidden in the formulation. A benevolent agent is pleased when someone has moved to a *higher level*, and is displeased when someone is *unhappy* or *not well off* or has moved to a *lower level*. Brandt does not define what these "levels" are, but it seems obvious that the moral aspect creeps in here: a benevolent person is pleased when something *good* happens to other people, whereas *bad* things happening to her fellow people make her feel discomfort. A formal definition of the morally acceptable appears not to suffice; we need a more substantive concept of the good for a moral theory.

[59] Brandt seems to notice the problematic nature of benevolence, namely, that it can implicitly include a moral aspect, for he treats a desire to be benevolent as a member of the same class as a desire to be moral for the sake of morality. A desire to be benevolent is a second-order desire which does not belong to the set of desires constituting the psychological point of view in Brandt's theory; see BRANDT 1979, 330.

[60] BRANDT 1979, 206–207, 215.

of moral rules. All deeds prohibited by these norms are forbidden, and anyone who (without a further reason) makes herself guilty of breaking such a moral rule, is morally reprehensible. All moral codes of fully rational moral agents include the prohibitions and obligations of this minimum code as their core. Depending on the amount of their benevolence the moral codes of fully rational agents also comprise other prohibitions and obligations the fulfilling of which enhances the well-being of the moral community. The norms exceeding the minimum code are, nevertheless, neither ethically compulsory, nor demanded. They simply bring extra moral good for the community and have, consequently, a supererogatory status.[61]

Brandt does not explicate the reason why all fully rational agents would accept at least the moral code of a perfectly selfish but fully rational agent, but the idea becomes intelligible when it is seen in the light of Brandt's utilitarianism. Fully rational moral agents, whether perfectly benevolent or selfish, are utilitarians.[62] The utility they want to maximize is welfare, determined by desires directed towards people's happiness and fulfilment of needs. The versions of utilitarianism, which fully rational moral agents tend to support, differ along the objects of moral concern and the degree of welfare of others. Perfectly selfish agents pay attention to others only to the degree this affects their own

---

[61] Being benevolent, or lacking this characteristic, marks the moral code of a rational person in the following way: to be a fully benevolent moral agent means that one prefers the options which maximize the net amount of happiness to be gained through time, no matter to whom this happiness is allotted. A fully benevolent agent is, nevertheless, not an altruist, but she grants herself and her own happiness the same weight as those of everybody else. BRANDT 1979, 215. A fully selfish moral agent, again, has no interest in the happiness of her neighbours. She is, however, not an egoist. To be an egoist namely involves supporting a moral code the central norm of which is that everybody must devote her life to the furthering of the egoist's well-being. Supporting that kind of a code is irrational: no one else but the egoist lets the code effect or regulate her behaviour. No one supports a moral code which only burdens her without ever bringing her any benefit. BRANDT 1979, 220–221. That it is theoretically possible to make this distinction between morally required and supererogatory actions, Brandt regards as the strength of his theory: there are many moral theories in which it is not possible to distinguish between obligatory and supererogatory norms. BRANDT 1979, 276–277, 289.

[62] BRANDT 1979, 208.

welfare. These agents accept the principle of mutuality: justice done to others for the sake of justice done to oneself. The principle of mutuality is also a minimum requirement for a socially functionable moral code: the costs of living according to the moral code do not exceed the benefits complience to it brings, and these benefits are great enough to justify the burden of its restrictions. Every fully rational moral person accepts the principle of mutuality as a cornerstone of morality. This secures a wider acceptance and adherence to the moral code, but it also makes the moral code more social: the best means of maximizing one's own welfare is to maximize the welfare of a group, or in a larger context, of a society.[63]

The difference between a benevolent and a non-benevolent fully rational agent can also be expressed by saying that they represent two different views of the role of moral objects as constituents of a morally significant situation. For a benevolent agent a situation becomes morally relevant as soon as there are sentient beings involved. A non-benevolent agent, by contrast, acknowledges the role of a moral object only indirectly: others bear the same relation to her as moral objects as she herself is a moral object in relation to others as moral subjects.

To conclude this section, I will take up some remarks related to the concept of a person explicated in the course of the study. Brandt's naturalist psychological theory explains, on the one hand, what is at issue in the institution of morality and what is, thus, morally relevant. On the other hand, the same theory is used as a normative criterion for the morally right, which in Brandt's theory equals the rationally correct. This is a typically utilitarian feature in the theory; the moral good is given a non-moral definition.

---

[63] Brandt 1979, 220–221. A perfectly benevolent, rational person also accepts the principle of mutuality, but she widens the scope of morality to include sentient beings the welfare, or the malfunctioning of which have no effect on her personal well-being. She does this because her morality is not just a means for securing her own good, but a guarantee for maximizing the welfare of a larger group of sentient beings, including, e.g., future generations,the mentally handicapped, children, foetuses, and animals; the scope of those involved depending on the degree of an agent's benevolence. Brandt 1979, 215–217, 221–222.

The concept of a person underlying the theory is that of a desiring agent. The desires and preferences of the agent form the moving force of all intentional action the aim of which is some form of desire fulfilment. What distinguishes moral action from other intentional action is that the aim of moral action is to produce the best net balance of desire satisfaction whereas non-moral action is not motivated by this aim. We could say of an agent performing these two types of actions that as a natural person the agent concentrates on enjoying the non-moral good of desire-satisfactions but that as a moral person she adopts the role of a rational producer of that good. Here the main features of the moral person also coincide with Brandt's definition of the morally relevant: what distinguishes a moral person from a natural one is just what makes something qualify morally — it passes the test of cognitive psychotherapy.

Furthermore, the moral person of Brandt's theory is a more or less benevolent agent who aims at maximizing the happiness of sentient beings by allowing a set of moral rules to guide her behaviour. The fact that some people are more benevolent than others does not constitute a moral difference between them, rather it is a natural difference, a variation in their psychological make-up. Benevolence is a natural characteristic which morality does not create but simply cultivates to a higher level.

The task of moral philosophy is to make people understand the nature of morality in the right light but when this has happened the non-moral point of view vanishes. The ideal moral agent, the fully rational person — a character we should all become like — regards everything from the moral point of view, because she has moulded and constantly moulds her desires, wants, and actions by cognitive psychotherapy. This justifies us in classifying Brandt's theory as a form of "moral imperialism".

1.2. R. M. Hare's universal prescriptivism

Hare calls his version of utilitarianism *universal prescriptivism*.[1] The core of his moral philosophy is the conceptual analysis of moral language. Knowledge of the specific logic of moral language offers a basis for a correct understanding of morality, upon which a normative moral system can then be built. In Hare's version of *normative ethics* we can distinguish a theoretical analysis concerning the structure of moral thinking, and a normative theory of practical moral reasoning. In the following, I will first examine Hare's theory of the two levels of practical moral reasoning, moving on to a presentation of his theory concerning the nature of moral language. Hare's normative moral theory will also be analyzed. In keeping with my thesis, these issues will be examined in order to explicate Hare's concept of a moral person.

1.2.1. THE TWO LEVELS OF MORAL REASONING

We cannot understand the nature of our moral reasoning correctly, Hare maintains, unless we notice that we employ two different kinds of thinking in our moral reasoning. This view is presented as an alternative to *intuitionism*, which Hare regards as a false moral theory. The mistake of intuitionism lies in the belief that we intuitively know what is morally right and wrong without any need or means to say what exactly it is that constitutes these moral properties. Intuitionists have been led to their

[1] See, e.g., HARE 1963, 16; HARE 1984, 228. I will mostly concentrate on Hare's *Moral Thinking. Its Levels, Method and Point*. (Clarendon Press, Oxford, 1984). Hare's *The Language of Morals*. (Oxford University Press, Oxford, 1952) and *Freedom and Reason*. (Oxford University Press, Oxford, 1963) are also referred to, especially to explain Hare's analysis of the meaning of moral terms. I treat Hare's theory as an unchanged whole without attending to its development; see HARE 1988, 201–205, where Hare gives his view about the unity of his work.

misconception by the fact that people have moral intuitions, and that they refer to them in their everyday reasoning. This does not, however, give us a reason to adopt intuitionism as a metaethical stance.[2] Instead, we have to notice that there are two main kinds of ethical thinking, which take place at what Hare calls the *critical* and the *intuitive level*.[3] In our everyday life we seldom think about ethical questions but act in a routine manner, or simply follow our moral intuitions. If we were asked to justify what we do, we would refer to our intuitions. This is quite adequate for most of the cases we encounter. At the intuitive level simple *prima facie* principles direct our actions, and we act mainly along the lines we have been taught. Accordingly, the intuitive level represents the conventional morality of our community as we have adopted it. The better the intuitions we have learned are, the better we act as moral agents.[4]

At the intuitive level we will realize that a moral decision must be made, and if the situation is not too complex or abnormal, we can solve it successfully by relying on what we feel is right or good.[5] There are, however, morally perplexing cases in which we notice that our intuitions do not guide us sufficiently. Applying norms in unfamiliar situations can

---

[2] "The most fundamental objection to the one-level account of moral thinking called intuitionism is that it offers no way of answering such a question [concerning conflicting intuitive prima facie rules]. The intuitive level of moral thinking certainly exists and is (humanly speaking) an essential part of the whole structure; but however well equipped we are with these relatively simple, prima facie, intuitive principles or dispositions, we are bound to find ourselves in situations in which they conflict and in which, therefore, some other, non-intuitive kind of thinking is called for, to resolve the conflict. [...] [intuitions] are not self-justifying; [...] To use intuition itself to answer such [justificatory] questions is a viciously circular procedure; if the dispositions formed by our upbringing are called into question, we cannot appeal to them to settle the question." Hare 1984, 39–40.

[3] Hare 1984, 40.

[4] "Because we are human beings and not angels we have adopted or inherited what I called the intuitive level of moral thinking with its prima facie principles, backed up by powerful moral feelings, and attached to rather general characteristics of actions and situations." Hare 1984, 59. Hare stresses the point that the intuitive prima facie principles are first, very general in their formulation, and second, overridable. The overridability is explained partly by the generality of the principles: no general principle can cover all the varying cases of human life. Prima facie principles must be held overridable also because there are many situations in which they conflict with each other. Hare 1984, 59–60.

[5] Hare 1984, 46–47.

be problematic, and we may face situations in which two or more moral norms, which we feel ethically binding, suggest mutually exclusive decisions. In such dilemmas our intuitions cannot really help us. What ought we to do? Now, it is these situations that call for rational decision-making and this can only take place at the so-called *critical level*.[6] The way to deal with such difficulties successfully is to abandon moral intuitions and to approach the acute problems by depending solely on the normal constituents of rational decision-making. At the critical level we can solve conflicts between intuitive prima facie principles, and formulate or redefine these principles so that they can be better applied to intuitive moral thinking.[7] In order to be able to reason in this more demanding and more complex way, we need to know what is really involved in our use of moral concepts.[8]

Hare's conception of the two levels of ethical thinking has, interestingly, a dual status. First, it is a *description* of the two levels of people's

[6] "What will settle the question [concerning conflicting intuitions] is a type of thinking which makes no appeal to intuitions other than linguistic. I stress that in this other kind of thinking, which I am calling *critical* thinking, no moral intuitions of substance can be appealed to. It proceeds in accordance with canons established by philosophical logic and thus based on linguistic intuitions only. To introduce substantial moral intuitions at the critical level would be to incorporate in critical thinking the very same weakness which it was designed to remedy. A philosopher will not be content with the intuitive props on which most moral philosophers rely, if he wishes his work to last." Hare 1984, 40. See also Hare 1984, 45, 49–50.

[7] "Critical thinking aims to select the best set of prima facie principles for use in intuitive thinking. It can also be employed when principles from the set conflict *per accidens*. Such employment may lead to the improvement of the principles themselves, but it need not; a principle may be overridden without being altered. [...] But besides the role of *selecting* prima facie principles, critical thinking has also the role of *resolving conflicts* between them." Hare 1984, 49–50. Hare is sure that all moral conflicts can be solved on the critical level, and that the very specific principles critical thinking produces for us are always overriding. Hare 1984, 59–60. Scanlon (1988, 133–135) criticizes Hare's model of the two levels of moral thinking for unnecessarily simplifying the reality of moral reasoning. Even relying on an intuitive moral principle as the basis of one's action in a particular case there is always a need for critical moral thinking.

[8] "At any rate, the first step that the moral philosopher has to take, in order to help us think better (i.e. more rationally) about moral questions, is to get to understand the meanings of the words used in asking them; and the second step, which follows directly from the first, is to give an account of the logical properties of the words, and thus of the canons of rational thinking about moral questions." Hare 1984, 4.

actual moral reasoning. Intuitionists, who do not understand this, give a false description of the nature of moral thinking. Second, Hare's theory establishes a procedure to be applied in morally perplexing situations: In non-problematic situations act according to your intuition, in a situation of conflicting norms or unclear intuitions, abandon your intuitions and think critically! Hare does not, however, pay attention to this double role of the two-level theory.

1.2.2. The logic of moral language

The study of the logic of moral terms[9] is, Hare maintains, essential for moral philosophy for it provides the only way to understand the conceptual nature of moral reasoning.[10] According to Hare's classification, the terms and statements of any language can be divided into three groups: 1) *pure descriptions*, 2) *prescriptions*, and 3) *value-statements*. Let us examine each of these groups respectively. The most important property of *descriptive* terms and statements is that they are *universalizable*. This means that when we know the semantic scope of a descriptive term we can use the term for describing all those objects which have such corresponding

---

[9] Hare does not make a distinction between a *term* and a *concept*, and a *statement* and a *proposition*, but simply speaks of terms and statements. It would be more appropriate, however, to use the terms *concept* and *proposition* instead, while the object of Hare's study is the *conceptual* nature of moral language.

[10] Hare presents his view as an alternative to ethical descriptivism, which according to Hare, offers a mistaken conceptual analysis of ethical questions. Ethical descriptivism, or naturalism as Hare also calls it, misses the target in its analysis by assuming that the task has been fulfilled after the linguistic conventions of some language group using moral terms have been laid bare. To do this is, however, far from sufficient for a moral philosopher, since defining the meaning of moral concepts is not just a matter of examining certain agreements concerning the use of moral terms. In contrast to the descriptivist view, Hare stresses that logical analysis of moral concepts involves most important decisions concerning substantive moral questions, because the logical properties of moral terms constrain the moral evaluations we can make. Hare 1963, 21; 1984, 6–7, 67, 70, 85.

properties.[11] Universalizability regulates the correct use of language so that one cannot, without contradicting oneself, state of two similar objects (or states of affairs) *a* and *b*, that *a* has a property *F* but that *b* does not have it.[12]

The second group of terms in Hare's analysis consists of *prescriptions*, by which he means ordinary singular imperatives. Imperatives cannot be reduced to indicatives: their use is determined by a different logic and by other kinds of conditions than those of indicative sentences. The most important difference is that whereas we may call indicative sentences true or false, this does not apply to imperatives. To assent to an indicative statement is to regard it as true, whereas assenting to an imperative sentence involves acting in accordance with the imperative. Similarly, to deny an indicative, is to call it false, but to deny an imperative is to refuse to act according to it.[13]

To summarize Hare's analysis of descriptive and prescriptive terms and statements, we can say that there are two pairs of conditions which regulate the logic of the indicative and the imperative mode respectively. First, it is self-contradictory to assent to an indicative statement and not to believe it to be true. Similarly, it is self-contradictory to yield to an

---

[11] HARE 1963, 9–10.

[12] "[...] in an apparently trivial, but at any rate unobjectionable, sense, any singular descriptive judgement is universalizable: viz. in the sense that it commits the speaker to the further proposition that anything exactly like the subject of the first judgement, or like it in the relevant respects, possesses the property attributed to it in the first judgement." HARE 1963, 12.

[13] "If we assent to a statement we are said to be sincere in our assent if and only if we believe that it is true (believe what the speaker has said). If, on the other hand, we assent to a second-person command addressed to ourselves, we are said to be sincere in our assent if and only if we do or resolve to do what the speaker has told us to do; if we do not do it but only resolve to do it later, then if, when the occasion arise for doing it, we do not do it, we are said to have changed our mind; we are no longer sticking to the assent which we previously expressed. It is a tautology to say that we cannot sincerely assent to a second-person command addressed to ourselves, and *at the same time* not perform it, if now is the occasion for performing it and it is in our (physical and psychological) power to do it. Thus we may characterize provisionally the difference between statements and commands by saying that, whereas sincerely assenting to the former involves *believing* something, sincerely assenting to the latter involves (on the appropriate occasion and if it is within our power), *doing* something." HARE 1952, 19–20.

imperative and not to act (now or in the future) in accordance with the command stated in the imperative. Second, an indicative conclusion may only be inferred from a set of indicative premises; and no imperative conclusion may be inferred from a set of premises which do not include at least one imperative premise. The last rule implies that no set of solely factual premises entitles us to deduce a norm from them (*Hume's guillotine*).[14]

In ethics, however, we mostly deal with so-called *value-statements*. Value-statements are characteristically both descriptive and prescriptive.[15] Consequently, all value-statements possess the logical properties of both indicative and imperative statements.[16] The descriptive meaning of value-statements implies that they must be *universalizable*.[17] Due to this logical property, we get a rule directing the accurate use of value-statements: if a value-statement *V* applies to some case *a*, then the same statement applies to every case *b*, *c*, *d*, etc., which possess the same relevant descriptive features as the original case *a*.[18] In addition to the characteristics of descriptive terms, value-statements also include a prescriptive element. Accordingly, their correct use requires that the logical rules of the imperative mode are followed. Thus, to designate something as, say, good includes an imperative: "This is good, so do it; or, like it". Calling something good implies that the speaker recommends it to others, and thus expresses a positive attitude towards the object in ques-

[14] Hare 1952, 28.

[15] Hare 1963, 4–5.

[16] Hare 1963, 26–27; 1984, 22.

[17] Hare 1952, 129; 1963, 21.

[18] "Now since it is the purpose of the word 'good' and other value-words to be used for teaching standards, their logic is in accord with this purpose. [...] The reason why I cannot apply the word 'good' to one picture, if I refuse to apply it to another picture which I agree to be in all other respects exactly similar, is that by doing this I should be defeating the purpose for which the word is designed. I should be commending one object, and so purporting to teach my hearers one standard, while in the same breath refusing to commend a similar object, and so undoing the lesson just imparted. By seeking to impart two inconsistent standards, I should be imparting no standard at all. The effect of such an utterance is similar to that of a contradiction; for in a contradiction I say two inconsistent things, and so the effect is that the hearer does not know what I am trying to say." Hare 1952, 134.

tion.[19]

There are four kinds of value-statements, depending on whether they are moral or nonmoral, or primarily or secondarily evaluative.[20] In primarily evaluative statements, or terms, the prescriptive component overrides the descriptive element, whereas in the secondarily evaluative terms the prescriptive element is secondary to the descriptive.[21] If we describe something with a secondarily evaluative term, for example, "industrious", it is still meaningful to ask whether it is good, or beautiful, or whether we should prefer, or like it.[22] This is not the case with primarily evaluative terms and statements; they are all both prescriptive and universalizable, irrespective of whether their content is moral or, say, aesthetic. The crucial difference between moral and non-moral primarily evaluative statements is that despite the universalizability of an aesthetic statement we can ignore it, i.e., we can let it not affect our action, whereas a moral statement which is primarily evaluative cannot be over-

---

[19] "I have said that the primary function of the word 'good' is to commend. [...] When we commend or condemn anything, it is always in order, at least indirectly, to guide choices, our own or other people's, now or in the future." HARE 1952, 127. "[...] in saying that it is proper to call a certain kind of man good (for example a man who feeds his children, does not beat his wife, &c. [sic.!]) we are not just explaining the meaning of a word; it is not mere verbal instruction that we are giving, but something more: *moral* instruction. In learning that, of all kinds of man, *this* kind can be called good, our hearer will be learning something synthetic, a moral principle. It will be synthetic because of the added prescriptiveness of the word 'good'; in learning it, he will be learning, not merely to use a word in a certain way, but to commend, or prescribe for imitation, a certain kind of man. A man who whole-heartedly accepts such a rule is likely to *live*, not merely *talk*, differently from one who does not. Our descriptive meaning-rule has thus turned into a synthetic moral principle." HARE 1963, 23. See also HARE 1952, 91, 117–118. Hare's interpretation of the logic of value-statements, bears a close affinity to that of the emotivists, as has often been noticed. See, e.g., WARNOCK 1981, 30. Hare himself strongly criticizes the emotivists and does not pay attention to this similarity. The fact that the prescriptive meaning does not exhaust the meaning of moral terms in Hare's theory does not alter the fact that this feature of his interpretation of the logic of moral terms still resembles that of the emotivists. See, e.g., HARE 1952, 144.

[20] 'Good' is a primarily evaluative moral term, whereas 'beautiful', although primarily evaluative, is a non-moral term. 'Brave' and 'artistic' are secondarily evaluative, the first a moral and the second a non-moral term.

[21] HARE 1952, 118; 1963, 24.

[22] HARE 1984, 17–18.

ridden: in virtue of its (logical) nature it demands to be paid attention to, and to be acted upon. If a moral and an aesthetic principle contradict each other it is part of the logic of these notions that the moral principle will always get prioritized.[23] Thus, the universalizability of moral principles means that they are universally binding.

The division between moral and non-moral evaluative statements plays a central role in Hare's theory — it demarcates the distinction between moral and non-moral value statements. Ethics can be distinguished from aesthetics on the ground that ethical statements cannot be overridden; primarily evaluative aesthetic statement can be disregarded at will, whereas moral statements must always be attended to. The criterion which distinguishes the two kinds of statements is that something cannot be overridden by other considerations when it concerns *other people's interests*: in ethics we deal with other people's interests, whereas aesthetic questions concern only the sensory and affective qualities of the aesthetic object.[24]

Hare's distinction between moral and aesthetic statements includes a definition of the *morally relevant*, although Hare does not discuss it. He takes it for granted that moral statements are expressed in primarily evaluative language and that they concern other people's interests. The language of the primarily evaluative binds morality to intentional action; the conditions for the correct use of this language are given in terms of action and abstaining from action. These are the minimal conditions which constitute a morally relevant situation. This definition also gives a rudimentary suggestion of the moral person in Hare's theory: an agent becomes a moral agent, or a moral subject when her actions have an effect on what is in other people's interests, and a person becomes a moral object when her interests are affected by what someone else does. Before examining Hare's view further, a critical comment is in order.

The difference between moral and aesthetic statements is, according to Hare's view, that moral statements concern other people's interests,

---

[23] Hare 1963, 168–169; 1984, 54.

[24] Hare 1984, 54.

whereas aesthetic statements do not. It is, however, possible that aesthetic matters do in fact affect other people's interests deeply.[25] Further, Hare's theory does not give any unarbitrary criteria for deciding whether something does or does not affect someone's interests. Consequently, one can admit the universal nature of some alleged moral principle but say that this principle does not represent a demand incapable of being overridden: one can treat it as a universalizable principle, but only of aesthetic nature. Hare's theory presupposes that moral terms, by their logical nature, demand to be acted upon. Despite this, the theory fails to determine non-disputable criteria for differentiating between moral and non-moral value-statements. Thus, Hare seems to assume that the *moral* nature of certain value-statements is (somehow) self-evident. In this respect he resembles the intuitionists he criticizes: they, too, presume that the special nature of moral statements is obvious, so that no criteria need to be explicated for distinguishing them from other kinds of statements. Hare could, however, answer this criticism by maintaining that as he only explains the logic of moral notions, it is not necessary for him to formulate any such criteria.

To clarify Hare's analysis of the logic of value-statements further, let us take an example in order to explicate the logic of primarily evaluative moral terms; I present his analysis of the term 'good'. To know the descriptive meaning of good involves being aware of the standards the speaker uses in her evaluation. The prescriptive meaning of good contains a befall, an estimation, or a prescription. This part of the meaning aims at directing the behaviour of the listener. Two questions can be asked in relation to the prescriptive meaning of good: First we can ask "Good what?"; and the answer then determines the class of objects within which comparisons are being made. The second question is: "What makes somebody call this good?", or: "Which qualities make the

[25] GRIFFIN (1990, 82) offers an example of this category: in France several car accidents occur yearly since drivers pay insufficient attention to their driving when travelling on country roads lined by beautiful trees. In spite of this, the trees are not cut down but are preserved for aesthetic reasons. Sticking to aestheticism in this case, nevertheless, affects other people's interests. Furthermore, by refusing to cut the trees the officials prefer aesthetic values to the moral ones which set out a norm to protect human life.

object good?". These questions remain similar irrespective of whether they concern the instrumental, or the moral good.[26] The evaluative meaning of the word remains, in its essential features, intact throughout the wide variety of the uses of the term.[27] The following example reveals the logic of the evaluatively used 'good': of three, equally expensive calculators, *a*, *b*, and *c*, *x* regards *a* as the best, but she chooses *c*. If *x* is sane in mind, we can say that she has not understood what "best" means: "best" refers to the object of some class we would choose before anything else.[28]

Deontically used obligatives (ought, must) belong to the same class of value-terms as "good". Thus, their correct use involves paying attention to the logic of both descriptive and prescriptive terms: sentences including a deontic obligative are universalizable, and follow the logic of imperatives. Hence, it is logically contradictory to say: "*x* must do *b* in a situation *c*, but if *y* is in the same (kind of) situation *d* she does not have to perform the action *b*." That this is true exemplifies the fact that we may not utter two dissimilar moral statements of two situations similar in their universal descriptive features without being guilty of a logical mistake.[29] This is, according to Hare, one of the most important features of moral language. The special logic of imperatives presupposes that when one uses sentences including a deontic obligative one may never break the following two rules:

> 1) from an indicative set of premises no conclusion including a moral term may be inferred;[30]

---

[26] "We thus have to distinguish two questions that can always be asked in elucidation of a judgement containing the word 'good'. Suppose that someone says 'That is a good one'. We can then always ask (1) 'Good what — sports car or family car or taxi or example to quote in a logic-book?' Or we can ask (2) 'What makes you call it good?' To ask the first question is to ask for the class within which evaluative comparisons are being made. Let us call it the class of comparison. To ask the second question is to ask for the virtues or 'good-making characteristics'." Hare 1952, 133.

[27] Hare 1952, 139–140, 143–144.

[28] Hare 1952, 103, 105–106.

> 2) a person cannot, without making a logical mistake, accept a principle which includes a deontic obligative, without following this principle herself.[31]

The logic of moral terms reveals two essential features of morality: first, it is *prescriptive* of nature. Second, morality attempts to answer the question: "What ought I to do?" with a prescription, and this prescription, for it to be valid, must be *universalizable*. These two aspects determine the correct use of moral language and Hare uses them to provide formal criteria for moral validity: the study of moral questions and the answers given to ethical problems must follow the logical rules set out by the conceptual analysis of moral terms.[32]

Hare's conception of morality is formal. He tries to develop a general frame of a universalizable moral code which is not bound to any specific

---

[29] "[...] to guide choices or actions, a moral judgement has to be such that if a person assents to it, he must assent to some imperative sentence derivable from it; in other words, if a person does not assent to some such imperative sentence, that is knock-down evidence that he does not assent to the moral judgement in an evaluative sense [...] This is true by my definition of the word evaluative. But to say this is to say that if he professes to assent to the moral judgement, but does not assent to the imperative, he must have misunderstood the moral judgement (by taking it to be non-evaluative, though the speaker intended it to be evaluative). We are therefore clearly entitled to say that the moral judgement entails the imperative; for to say that one judgement entails another is simply to say that you cannot assent to the first and dissent from the second unless you have misunderstood one or the other; and this 'cannot' is a logical 'cannot' — if someone assents to the first and not to the second, this is in itself a sufficient criterion for saying that he has misunderstood the meaning of one or the other. Thus to say that moral judgements guide actions, and to say that they entail imperatives, comes to much the same thing." HARE 1952, 171–172. See also HARE 1963, 29.

[30] This principle is inferred from the rule governing the logic of imperatives and Hare's analysis of moral terms. The principle directing the use of imperative sentences is as follows: "No imperative conclusion can be validly drawn from a set of premisses which does not contain at least one imperative." HARE 1952, 28.

[31] "If a person says 'I ought to act in a certain way, but nobody else ought to act in that way in relevantly similar circumstances', then, on my thesis, he is abusing the word 'ought'; he is implicitly contradicting himself. [...] What the thesis does forbid us to do is to make different moral judgements about actions which we admit to be exactly or relevantly similar. The thesis tells us that this is to make two logically inconsistent judgements." HARE 1963, 32–33. See also HARE 1984, 115.

[32] HARE 1984, 6–7, 21.

normative morality. Additionally, this frame must be applicable to substantive moral questions. What significance does it have for a person's morality that she is aware of the logical properties of moral language? Does knowledge concerning the nature of moral language have *moral* relevance? If it does, do we then have a moral obligation to get acquainted with the nature of moral language? This is a second-order moral obligation which concerns, not any moral question as such, but our awareness of the nature of such questions. Hare seems to think that examining the nature of moral language makes a difference for the kind of morality we practice, but he does not offer any grounds for this belief.

The role of the conceptual analysis of moral terms in Hare's theory gives us a reason to suspect that Hare does not pay enough attention to the fact that the logically correct understanding of the use of moral terms and statements, and people's actual adherence to moral rules are two separate issues; for him they seem largely to be the same thing.[33] Thus, Hare appears to rely on the force of logical correctness as a guarantee for people's following moral prescriptions.[34] An implicit premise is lodged in this: if we understand the logic of moral terms and statements, we do not offend against morality, for offending against morality means offending against logic, and no rational person will ever do that. It is still doubtful whether the logic of moral terms has the same status as a criterion for rational action as the logic of purely indicative statements. Whereas no one can be considered rational if they disregard the rule of contradiction in their thinking, it is not impossible to imagine a person who, even after having understood Hare's logical analysis of moral terms, reacts with the comment "So what?" to an incorrect use of moral terms, and who in her action transgresses against the rule of moral universaliza-

[33] Hare deals with this question as he discusses weakness of will (Hare 1984, 57–60). According to his interpretation weakness of will can be best interpreted as an analogy with a conflict between prima facie rules on the intuitive level: it is an example of conflicting prescriptions. I have to choose whether I offend against a moral rule or whether I "disappoint my own appetites". When examined on the critical level, the problem dissappears. Hare 1984, 60. Nagel (1988, 102–112) claims that the elements in the logic of the moral concepts cannot in fact constitute Hare's form of utilitarianism by virtue of the logical properties of these concepts. The role these concepts have in Hare's theory is that of moral claims presented without a justificatory foundation.

bility. We can call such person immoral but not necessarily irrational.[35]

We have now explicated the theoretical structure of Hare's moral theory. The specificity of the moral lies in the special logic of moral language which makes ethical statements both universalizable and prescriptive. Furthermore, moral statements differ from other universalizable and prescriptive statements because they concerns other people's interests. Hare's analysis of moral language implies that morality is part of the intentional, and a question becomes morally relevant when other people's interests are affected by what the agent does. Hare does not

[34] Hare stresses the point that the conceptual analysis of moral terms does not provide us with any substantive moral principles: "Offences against the thesis of universalizability are logical, not moral. If a person says 'I ought to act in a certain way, but nobody else ought to act in that way in relevantly similar circumstances', then, on my thesis, he is abusing the word 'ought'; he is implicitly contradicting himself. But the logical offence here lies in the *conjunction* of two moral judgements, not in either one of them by itself. The thesis of universalizability does not render self-contradictory any single, logically simple, moral judgement, or even moral principle, which is not already self-contradictory without the thesis; all it does is to force people to choose between judgements which cannot both be asserted without self-contradiction. And so no moral judgement or principle of substance follows from the thesis alone. [...] What the thesis does forbid us to do is to make different moral judgements about actions which we admit to be exactly or relevantly similar. The thesis tells us that this is to make two logically inconsistent judgements." HARE 1963, 32–33. In HARE 1984 there seems to be a new interpretation (although Hare denies this, see HARE 1984, 6) of the relationship between the conceptual analysis and the normative theory: Hare maintains that the prescriptivity of moral terms provides the grounds for their practical bindingness: "[...] moral words have [...] a commendatory or condemnatory or in general prescriptive force which ordinary descriptive words lack. The person who thinks that the fact that an act would be wrong is no reason at all for not doing it shows thereby that he has not fully grasped the meaning of the word." HARE 1984, 71. And: "[...] I shall be maintaining that, if we assumed a perfect command of logic and of the facts, they would constrain so severely the moral evaluations that we can make, that in practice we would be bound all to agree to the same ones. [...] the freedom which we have as moral thinkers is a freedom to reason, i.e. to make rational moral evaluations; and the rules of this reasoning, which are determined by the concepts occurring in the questions we are answering, bring it about that, over the most important part of morality, we shall, if we are rational, exercise our freedom in only one way." HARE 1984, 6–7.

[35] In this respect Hare radically differs from Mackie who, despite acknowledging the formal validity of the universalizability thesis, nevertheless dismisses it as a basis for any substantial moral principle. According to him, we can whenever we want to, both affirm the formal plausibility of a moral principle and disregard it as affecting our action by placing ourselves outside the moral realm; see MACKIE 1990, 85–88. For a more detailed criticism on this point see also SINGER 1988, 147–159.

explicitly deal with intentional actions, but it seems that the underlying theory of action resembles that of Brandt's. An agent's desires are the moving force of her action and she can control her choice of means for achieving the desired goal by her reason. The general scheme of intentionality becomes morally relevant through a moral object, i.e., another person who has autonomous interests. This means that a person cannot have moral duties towards herself; she does not become the moral object of her own action. But before continuing the analysis of Hare's concept of a moral person, we must examine his normative moral theory.

1.2.3. The premises of moral thinking

In moral thinking, we need to ponder what is essential in situations in which a moral decision is called for? On which premises do we base our moral reasoning? Any morally significant situation involves *people* and their interests or *preferences*. Moral reasoning, which leads to a particular outcome in a concrete situation, consists in answering such questions as: who is involved? What are their preferences? Which actions can be regarded as viable alternatives? What effects do these actions have on people and on their desires or preferences?[36] For Hare, finding an answer to these questions is essentially what substantive moral reasoning is about.

The choice of the facts on which moral reasoning is based also has significance in relation to Hare's concept of a person. According to Hare, desires or preferences move a person to action. When a person deliberates over what she is to do, she has to consider which preferences her actions might affect. If her decision only concerns herself, she may choose what to do on purely prudential grounds. In this case we can say that the agent is a natural person whose action is only motivated by her

[36] Hare 1984, 89, 90–91.

preferences. But if the agent finds herself in a situation in which the interests (i.e., the fulfilment or deprivation of desires) of several people are concerned it can be conceived of as a morally relevant case which makes the agent a moral person. People's preferences define the morally relevant situation but they also define the basic concept of a moral person in Hare's normative theory. The fact that a person has preferences and that her action is directed in line with them defines the most important characteristics of the person both as a moral subject and as a moral object. A person becomes a moral object when someone else's actions affect her preferences, and she shifts in the role of a moral subject when what she does has an effect on the preferences of others.

Apart from what has been said so far, preferences also employ another role in Hare's theory: they build the bridge between facts and norms. The logical nature of preferences is such that a preference — somebody's desiring/preferring something — always includes a prescription directed at others: "Satisfy my preference!".[37] Consequently, when we analyze a morally perplexing situation we actually list the preferences of different people, which embrace prescriptions that follow the logic of imperatives. As premises of moral reasoning, preferences thus include imperative clauses entitling us to make valid imperative inferences from them. More explicitly, we can deduce norms from facts, that is, moral standards from people's preferences. Hare's analysis of desires as prescriptions directed towards others is a crucial building block in his theory, connecting the logic of moral terms, or the meta-ethics with normative ethics; but Hare pays astonishingly little attention to this part of his argument.[38]

If we proceed with our inquiry two further difficulties in Hare's analysis arise. As was noted earlier, people's preferences form the facts we

[37] "We shall see that the method of critical thinking which is imposed on us by the logical properties of the moral concepts requires us to pay attention to the satisfaction of the preferences of people (because moral judgements are prescriptive, and to have a preference is to accept a prescription); and to pay attention equally to the equal preferences of all those affected (because moral principles have to be universal and therefore cannot pick out individuals)." HARE 1984, 91.

[38] For a similar remark, see SEANOR & FOTION 1988, 4.

have to consider in a morally relevant situation. Now, it seems legitimate to ask whether we are ever morally required to perform an action no one who is involved in the situation prefers? I will show that both a negative and an affirmative answer to this question reveal a defect in Hare's theory. First, if we are never to act in a way which ever contradicts anyone's preferences, then we must seek out the morally viable alternatives among those which square with the preferences of the people involved. If this were the case, people's preferences would determine all (formally) acceptable ethical obligatives in a given situation.[39] Promoting this alternative has odd consequences, however; if the preferences of those involved determine the range of our possible moral decisions, and thus what is ethically right and wrong, the basic moral concepts will become arbitrarily dependent on the emotional and psychological constitution of each affected party. We cannot determine moral correctness independently of particular situations with particular parties; and we cannot make a morally justifiable decision unless we know what the preferences of the people involved are. Hare's way of analyzing the moral constituents of a situation focuses attention not on what people do, but on what they desire. To take an example: it would be justified for someone to commit any crime provided that this did not infringe anyone's preferences. But this is difficult to accept. Hare can avoid the difficulty only by maintaining that people's preferences are uniform up to the degree that all people would always want largely the same outcome, depending only on the position they occupy in a given situation.[40]

---

[39] Hare's remark (HARE 1984, 54) concerning the criteria for recognizing a morally significant situation from a non-moral one gives support to this interpretation. Hare, namely, maintains that a situation where other people's *interests* are not affected is not a case requiring ethical judgement. There is even more material in favour of this interpretation: "We are committed by the formality of our method to a Benthamite answer to the basic question: equal preferences count equally, whatever their content. This is because the only question the method allows us to ask is, What shall we rationally universally prescribe, or from an impartial standpoint prefer, if we are fully informed and make no logical mistakes? There is no place here for discrimination, at the critical level, between pleasures or preferences because of their content." HARE 144–145. See HARSANYI 1988, 90–99 for a similar remark and his objection to the possibility of immoral preferences in Hare's theory.

Let us now examine the other alternative to the question whether people's preferences exhaust the morally viable courses of action in a given situation, taking the view that we are sometimes morally required to act in a way no one involved desires.[41] This conclusion means, however, that the logic of moral terms and people's preferences are not the only source of moral reasoning. Were this the correct interpretation, Hare's analysis of the logic of moral language and of the constituents of a morally significant situation would be incomplete. This would be a failure for Hare's metaethical project; facts and logic, as presented by Hare, do not exhaust the scope of moral alternatives and consequently are unable to provide the moral deliberator with the tools needed for ethical decision-making. Hare can escape this criticism by referring to his theory of the two levels of moral thinking. I will sketch a possible argument for Hare which serves to rescue his theory.

Hare might maintain that it is only on the critical level of moral thinking that we can determine the relationship between people's actual preferences in a given situation and the morally viable course of action. It is, nevertheless, an error to think that we apply this level very often; we seldom have to reason on the critical level, but mostly act in accordance with our adopted, traditional prima facie principles which help us to describe the moral situation with its morally crucial features. Thus, refer-

[40] What HARE (1984, 13–14) says gives credence to this assumption: "Suppose that we know enough about human beings to be sure that they have, for the most part, certain desires, aversions, and the like. And suppose that, given these desires, and given a certain hypothesis about what the moral words mean, and thus about what forms of moral reasoning are valid, and given, further, the assumption that human beings do, on the whole, with some exceptions, reason validly, we can derive from the hypothesis about meanings the prediction that human beings will reach certain moral conclusions in their reasoning, and thus come to have certain moral opinions. [...] In short, we are allowed to retain, provisionally, those hypotheses about the meanings of words which, if true, and if we are right about what human beings are like, would account for them having the moral opinions which they do have."

[41] Hare's text also lends support to this interpretation: "We are to assume, when we come to universalize our prescriptions, as morality demands, that we have to consider only those prescriptions and preferences of others which they would retain if they were always prudent in the sense just defined. Our knowledge of the facts, so far as we manage to emulate the archangel in attaining it, will enable us to say what others would prefer if they were prudent [...]" HARE 1984, 105–106.

ring to imaginary examples, which involve people with odd preferences, distorts the whole nature of ethical thinking. If we have received a good moral education we will keep to simple deontic rules. Thus, the whole problem — whether we are morally required to act in a way nobody involved in a given situation would desire — misses the point. At the intuitive level we would never come to pose this question.[42]

We can, however, still offer an objection: Hare's answer is plausible only in our world where we already have a relatively solid social moral institution. It constitutes the given for our ethical thinking from which our moral intuitions arise. Now Hare claims that, although the intuitive level is sufficient for normal ethical decision-making, we have to make use of the critical level when we are deciding which kind of rules to accept as prima facie principles, and what the order of application in the case of two, or more, conflicting rules is. But can we, just by applying critical thinking, arrive at the prima facie principles we now have? Would referring to facts and logic be enough for establishing the kind of good prima facie rules Hare regards as being so essential for sound morality?[43] I suggest that we could not. The role of preferences on the critical level is crucial, for they provide us with the normative variants for solving the situation. Consequently, it is not possible to infer a moral system solely on the basis of "logic and facts". Or better, it is not possible unless we presuppose that people's preferences are factually relatively similar irrespective of time and place. Therefore, Hare does not succeed in doing all the work he claims to have accomplished in his theory by just appealing to facts and logic, but that he, in fact, employs other, unexplicated premises as parts of his argument.

---

[42] Hare reasons along this line when arguing against "fanatics"; see HARE 1984, 171–172.

[43] Hare's theory presupposes that we make up two kinds of moral rules; first prima facie rules to be used on the intuitive level, and the non-overridable, very detailed rules of the critical level. HARE 1984, 69–60. The prima facie rules are the rules we must teach our children, and to which we refer when we list our moral commitments. The problem with this view is that it presupposes that we can construct the prima facie rules with the help of critical thinking. In critical thinking we can, nevertheless, rely only on people's preferences and the logical properties of moral language: it is difficult to see how we could deduce the prima facie rules actually directing people's behaviour from them alone.

1.2.4. The method of identification

The importance accorded to people's preferences and their satisfaction reveals the utilitarian character of Hare's theory: the satisfaction of preferences is the aim of morality. There are, however, different kinds of utilitarian theories, depending on the nature of the morally relevant preferences. For Hare, the preferences which merit our moral concern are people's *present desires*, as they happen to be at the time of moral decision-making. The only thing a moral deliberator must take into account is the intensity of the preferences; she must consider all preferences without discrimination on any qualitative, or procedural criteria.[44] There are three kinds of difficulties connected with this view, all of which Hare maintains his theory is capable of solving; first, how can we overcome the bias everyone tends to have in their own preferences in comparison to those of others; second, how can we solve the problem of desires varying with time; and finally, how can we exclude the desires which are morally intolerable, but which must be taken into account as part of the facts which make up a morally significant situation.

For the purpose of overcoming the bias in oneself Hare introduces the *method of identification*. The moral agent must consider the position of each party involved in the morally problematic situation by vividly imagining herself in the place of each one. This must be done, not simply by thinking what it would be for me to be in the place of the other, but what it would be for me if I actually were the other, with her preferences, psychological qualities, and so on.[45] Hare backs up his method of identification with a conceptual move; the logical structure of a sentence is not affected by a change in the personal pronoun, so we can, in the sentence "I suffer pain", replace the first pronoun singular with another personal pronoun, e.g., "you".[46] Thus, the formal conditions of moral language rule out consideration of the personal identity of the person concerned, allowing only for the present preferences, and the consequences different courses of action have for the satisfaction of these preferences.[47]

---

[44] Hare 1984, 144–146.

Despite the moral unimportance of personal identity the concept still has some relevance for Hare's theory. Hare briefly considers, but quickly dismisses, the view according to which we could compare preferences without the concept of personal identity. He describes the special relationship people have with themselves by saying that the word "I" is not only descriptive in its meaning, but partly prescriptive, too. By this he means that identifying oneself with some person always involves identifying oneself also with the person's prescriptions.[48] Further, "by calling some person 'I', I express at least a considerably greater concern for the satisfaction of his preferences than for those of people whom I do not so designate."[49] This suggests that identifying oneself with a set of prescriptions gives us a criterion for determining personal identity: I am the

---

[45] "I do not wish to be taken as claiming that we can ever in fact have full knowledge of other people's experiences. [...] It would be wrong to claim that our imaginations somehow *inform* us of what the experiences of others are like. Imagination is a very common source of error; it can just as well be of experiences and preferences which they do not have as of those which they have. But if we do know what it is like to be the other person in that situation, we shall be (correctly) imagining having those experiences and preferences, in the sense of knowing or representing to ourselves what it would be like to have them; and this [...] involves having equal motivations with regard to possible similar situations, were we in them." HARE 1984, 95. "In general, when I say that somebody who would be in a certain situation would be *myself*, in so saying I express a concern for that person in that hypothetical situation which is normally greater than I feel for *other people* in the same situation. To recognize that that person would be myself is already to be prescribing that, other things being equal, the preferences and prescriptions of that person should be satisfied. This is what is involved in 'identifying' with that person." HARE 1984, 221. See also HARE 1984, 96–97, 98.

[46] "Actually the short answer to the problem about the *meaning* of statements about other people's states of mind is that terms like 'I' and 'you' have no *descriptive* content in the strict sense; that is to say, if you and I just changed places, the world would be no different in its universal properties. So the meaning of the *predicate* in 'You are in pain' is exactly the same as in 'I am in pain'[...]" HARE 1984, 123. See also HARE 1984, 120–121.

[47] The effect of this move is that interpersonal problems become intrapersonal, and consequently, comparing preferences, so central to Hare's utilitarianism, becomes simpler: "So we have in effect not an interpersonal conflict of preferences or prescriptions, but an intrapersonal one; both the conflicting preferences are mine. I shall therefore deal with the conflict in exactly the same way as with that between two original preferences of my own." HARE 1984, 110. See also HARE 1984, 128.

[48] HARE 1984, 96–97.

[49] HARE 1984, 98.

person whose set of prescriptions I would most intensively hope to get satisfied. Hare wants to avoid the discussion of personal identity; therefore we cannot decide what would establish full personal identity in his theory, but I suggest that the criterion just explicated is a necessary condition for determining it. It also remains unclear what kind of significance personal identity can have for morality because we are not allowed to pay attention to the individual whose preferences are in question in our ethical decision-making.[50]

Hare's solution to the problem of preferences which change with time is analogous to his answer concerning the preferences of different people. While the identity of distinct persons is not important in the theory where several people are concerned, neither does it matter in the case of the temporarily distinct and mutually incompatible desires of a singular person. We have to consider the future desires, whether our own or those of others, as if they were our present desires.[51]

The third difficulty Hare's utilitarian theory is likely to encounter is the problem of *evil desires*, or *immoral preferences*. All preferences, whether those of the Marquis de Sade or Mother Teresa, must be given equal attention. A preference, and the prescription included in it, qualifies for

[50] RICHARDS [1988, 119–120] criticizes Hare for confusing "the central ethical imperative, treating persons as equals, with the very different idea that preferences or desires should be treated equally". This is a sign that Hare's universal prescriptivism, as a form of utilitarianism, fails to do justice to the separateness of persons which is a fundamental fact of ethics.

[51] "There is thus no possibility of discounting the future, because we have to imagine it as present. This is a necessity for critical moral thinking, because [...] the universalizability of moral judgements puts a prohibition on the occurrence of time-references in moral principles; mere dates, by themselves, cannot have moral relevance." HARE 1984, 101. "Even in the case of experiences that we have actually had, and which we should surely be able to compare in this way if we can compare anything, it looks as if it can be done only by reducing past preferences to present ones. We imaginatively suppose that we could have the choice of having one of these experiences or the other, and form a preference *now*." Hare 1984, 125. There is, however, a line of thought which points in another direction in Hare's theory and which would, if elaborated further, involve a change in Hare's version to solve the problem of changing preferences. HARE (1984, 104–106) refers to Brandt (BRANDT 1979) and presents Brandt's idea for determining the rationality of one's preferences. According to this suggestion, we are only to consider our present desires, and to act along the ones that are rational in the light of our present knowledge and reasoning.

directing moral action if it passes the formal tests constituted by the logical rules regulating the use of moral concepts.[52] Hence, an agent may only accept a preference if the consequences of the action that accords with the preference are such that the agent could accept them were she in the position of any of the parties concerned. When evaluating the preferences, the agent must apply the method of identification. She must reflect on the possible courses of action and their outcomes for different people by identifying herself with each of the parties involved. Now, Hare maintains, that the test of identification shows that no one could accept the consequences of the preferences of, say, the Marquis de Sade for herself, were she in the position of the victim. This means that such preferences have failed the formal test of universalization, and cannot thus be accepted as norms guiding action.[53]

It is not, however, clear that the formal criteria for disqualifying the immoral preferences have the effect Hare claims them to have. The crucial point in his argument is the presupposition that no one would choose to be an object of deeds we normally regard as immoral. Hare does not, however, present any "proofs" for there being no such people.[54] We can at least think of someone whose preferences were, in our eyes, strange and abhorrent, but who would not have anything against being in the place of an object for someone with her preferences; to

[52] HARE 1984, 141, 144–145.

[53] "We retain, all of us, the freedom to prefer whatever we prefer, subject to the constraint that we have, *ceteris paribus*, to prefer that, were we in others' exact positions, that should happen which *they* prefer should happen. The requirement of universalizability then demands that we adjust these preferences to accommodate the hypothetical preferences generated by this constraint, as if they were not hypothetical but for actual cases; and thus, each of us, arrive at a universal prescription which represents our total impartial preference (i.e. it is that principle which we prefer, all in all, should be applied in situations like this regardless of what position we occupy.) What has happened is that the logical constraints have, between them, compelled us, if we are to arrive at a moral judgement about the case, to coordinate our individual preferences into a total preference which is impartial between us. The claim is that this impartial preference will be the same for all, and will be utilitarian." HARE 1984, 226–227.

[54] See, e.g., HARE's (1988, 245–246) answer to the criticism that his theory creates space for immoral preferences. Hare seems to take it for granted that examples depicting people with cruel and immoral desires are taken from the world of phantasy, not from real life.

claim that there are no such people is a factual, not a conceptual statement.[55] Hare's test for discerning immoral preferences stands and falls with the presupposition that people's preferences largely coincide.

To conclude, the Harean person is an intentional agent whose aim is to maximize the satisfaction of preferences. The preferences she aims at maximizing when no one else is involved are her own strongest preferences, and in doing this she is not a moral person but simply an intentional agent, or a natural person, who acts rationally. When several people's interests are affected she moves into the moral realm. As a moral object the fulfilment of her own interests depends on what others do; as a moral subject she must aim at maximizing the strongest preference of the situation, on condition that the rule proscribing the action can be universalized to apply to all parties. This means that preferences — which define both the natural and the moral person in Hare's theory — also play a crucial role both in explaining intentional action and in defining the morally relevant.

For Hare, it is a morally neutral, natural fact that people have preferences and that they aim to fulfil them by their action. This reveals the utilitarian nature of his theory. The starting point of morality is non-moral and natural, and becomes moral through a specific qualification. It is, however, noteworthy that Hare's theory includes a strong presupposition that people's preferences are uniform, and that no one would want to be the object of deeds we normally consider as immoral. We need this presupposition if we are to get the moral system Hare presupposes we can derive from the premises he grants us. In addition to the central role preferences have in a moral situation, we also have to make use of the specific logic of moral language in our ethical deliberation. Accordingly, we may only choose such actions which conform with the rules of moral language. Whether we are morally obliged to do this, remains however, controversial.

---

[55] In defence against such examples Hare simply states: "It is fairly obvious that no real case like this [somebody greatly enjoying another's suffering] will occur." HARE 1984, 141.

1.3. "The true view of ourselves" — Derek Parfit's theory

1.3.1. The aim of moral philosophy

Derek Parfit's, *Reasons and Persons*[1] is one of the most discussed philosophical books of recent years. It is particularly well-known for its treatment of personal identity,[2] but Parfit's other themes, including morality and rationality, as well as the moral status of future generations, have also received considerable attention.[3] In fact, the book consists of four comparatively independent works[4] which nevertheless all serve Parfit's main goal — the attempt to construct a new and, as he suggests, true basis for ethics.[5] Parfit repudiates almost all previous moral theories, especially those tied up with a theistic worldview. According to him, "the history of ethics is just beginning".[6] In the following I will concentrate mainly on examining how Parfit evolves and argues for his moral theory. I pay special attention to his discussion concerning the connection between the nature of personal identity and his utilitarian moral theory.[7]

The main difficulty with Parfit's book is that he leaves most of his

---

[1] Derek Parfit, *Reasons and Persons.* Clarendon Press, Oxford, 1984.

[2] Before *Reasons and Persons* Parfit was already known as the writer of several articles on personal identity, e.g.: "Personal Identity", *Philosophical Review* 80, 1970; "On 'The Importance of Self-identity", *The Journal of Philosophy* 68, 1971; "Personal Identity and Rationality", *Synthese* 53, 1982.

[3] *Reasons and Persons* has been widely reviewed and commented upon: see e.g. Sidney Shoemaker, "Reasons and Persons." *Mind* 1985, 443–453; Grant Gillet, "Brain Bisection and Personal Identity." *Mind* 1986, 224–229; Ursula Wolf, "Was es heißt sein Leben zu leben." *Philosophische Rundschau* 1986, 242–265; Nathan L. Oaklander, "Parfit, Circularity, and the Unity of Consciousness." *Mind* 1987, 525–529; David Cockburn, "Critical Notice." *Philosophical Investigations* 10, 1987, 54–72; Don Locke, "The Parfit Population Problem." *Philosophy*, 1987. 131–157; T.L.S. Sprigge, "Personal and Impersonal Identity." *Mind* 1988, 29–49. Parfit's ideas have also aroused a lively debate in literature, see e.g. Carruthers 1986, Gillet 1987, Trupp 1987, Glover 1988. Comments about Parfit's ideas vary greatly, for some (Glover, Shoemaker, Trupp) he is one of the most ingenious thinkers of our time, whereas for others (especially for Cockburn, Oaklander, and Wolf) his ideas seem totally misguided.

central concepts undefined, and attaches many, it seems, mutually inconsistent meanings to them.[8] To give but one example, Parfit does not define what he means by *morality*. He often refers either to "morality", or to "common sense morality", which he calls *M*. The way in which Parfit applies these terms implies that morality and common sense morality *M* mean for him the same thing. His use of the concepts reveals also that *M* consists of a collection of deontological moral principles inhibiting killing, stealing, and the like. Additionally, *M* tells us also to be disposed in a certain manner, i.e., to be biased towards people to whom we have a spe-

[4] Part one, 'Self-defeating theories' deals with three, according to Parfit's judgement, generally accepted theories that provide people with reasons for acting. Parfit tries to show that, unless modified, all these theories are in fact self-defeating and, therefore, incapable of serving as a basis for consistent action. Part two 'Rationality and time' is a discussion about the commonly adopted bias towards the future and of its irrationality. In this part Parfit attempts to refute what he calls the *Self-interest theory of rationality*, or '*S*'. Part three 'Personal identity' is an attempt to show how we are not what we think we are, and how this mistaken view distorts the way people conceive of themselves and of the nature of reality. Part four 'Future generations' tries to find a justification for caring and taking responsibility for the future generations and their well-being. Parfit admits to having failed in this part, the argument falling short of the theory he is looking for. The list of contents of the book, pages xi–xv, provides a crude summary of Parfit's lines of thought.

[5] Parfit refers to himself as a revisionist, Atheist moral philosopher. Parfit 1984, x, 453–454.

[6] It seems that moral theories bound up together with, or based on religious beliefs, have no value in Parfit's eyes. Because moral philosophy has been made a life's work by atheists only since the 1960's there has not been enough time for this field of knowledge to flourish, so: "Compared with the other sciences, Non-Religious Ethics is the youngest and the least advanced", and: "Belief in God, or in many gods, prevented the free development of moral reasoning. Disbelief in God, openly admitted by a majority, is a very recent event, not yet completed. Because this event is so recent, Non-Religious Ethics is at a very early stage. We cannot yet predict whether, as in Mathematics, we will all reach agreement. Since we cannot know how Ethics will develop, it is not irrational to have high hopes." Parfit 1984, 453–454.

[7] In part two of the book Parfit presents a "duel" between his revised theory of morality and rationality, *CP*, against the self-interest theory *S*. The section contains many interesting topics, but also severe mistakes and muddles, and should be treated with more precision than is possible in this connection. I will, however, refer to the form of Parfit's argument whenever it is necessary for the purpose of our present theme.

[8] This he does because: "[...] it [defining the central concepts] would take at least a book to give a helpful explanation, I shall waste no time in doing less than this." Parfit 1984, ix.

cial relation: our near relatives and our own pupils or patients. Parfit calls these connections *M-relations*.[9] Parfit also claims that there are many propositions which people universally or widely promote, without giving any evidence to support this opinion.[10]

Parfit's starting point is the *acting agent*. He maintains that as acting agents people want to know what they have most reason to do. There are several different kinds of theories offered in answer to this question; some of them are moral theories, others theories about rationality.[11] According to Parfit, people previously thought it rational to be moral. This conviction was founded on theistic belief. The reasoning in support of this view went along the following lines: every person has most reason to act according to her own interest; this is what Parfit calls the *self-interest theory of rationality* or *S*.[12] Furthermore, it is in everyone's interest that the almighty God upon whom everyone's eternal destiny depends accepts their deeds. God disapproves of actions which are morally bad and punishes those whose lives have been reprehensible with the horrors of hell. Thus, it is in everyone's interest and, therefore, rational for everyone to act as morality requires.[13]

---

[9] Parfit 1984, 102. Because the concepts have been left undefined, one soon begins to suspect that Parfit changes the meaning of his central terms arbitrarily as his reasoning proceeds, to enable him to reach the conclusions he is aiming. This is particularly evident in part two of the book where the alternative theory for *S* — *CP* — changes its outlook and structure in a quite unruly fashion. This gives a reason for Cockburn to state that "the whole discussion [...] (in Sections 53–61) is seriously confused." Cockburn 1987, 59.

[10] Parfit uses the term *Common-Sense morality*, or *M*, for referring to a view on life he assumes to be universal and uncontroversial. See, e.g., the discussion in section 36 (Parfit 1984, 95–98). In part one of his book he claims *M* to be directly self-defeating, but he actually never defines what he means by *M*. Another, astounding, example of taking things for granted is the self-evident way in which Parfit supposes the *Self-interest Theory* to have been "long [...] dominant in our intellectual tradition." Parfit 1984, 192.

[11] "Many of us want to know what we have most reason to do. Several theories answer this question. Some of these are moral theories; others are theories about rationality. When applied to some of our decisions, different theories give us different answers. We must then try to decide which is the best theory." Parfit 1984, 3.

[12] The core of this theory of rationality, which Parfit claims to be the most widely held, is that: "It insists that a rational agent give supreme weight to his own self-interest, *whatever* the costs to others. It insists that a rational agent must be biased in his own favour." Parfit 1984, 192.

Nowadays, Parfit claims, only a few people believe in God or in an afterlife. This has caused a dilemma for us, for most of us still think it rational to act in accordance with our own best interest, but it no longer seems rational to act morally. To explicate Parfit's thought we can say that the question concerning the connection between rationality and morality now appears in a different metaethical perspective than before. What we now consider to be rational, i.e., fulfilling our self-interest, no longer accords with what we regard to be moral. To solve the dilemma we should, it seems, either change our understanding of morality or find a new conception of rationality. In Parfit's words, the problem of modern moral philosophy would be settled if it were possible to show that rationality and morality coincide. For this purpose, a *unified theory* of rationality and morality should be constructed. Consequently, Parfit sees his main task as a moral philosopher to create this theory or at least to bridge some of the gaps dividing morality and rationality as they are understood currently.[14] He also suggests that his version of moral theory, *Critical Present-aim Theory CP*, is the best candidate for a unified theory. At the end of his work Parfit has to acknowledge that he has failed to accomplish the task he has set himself: his argument does not, even in his own estimation, succeed in proving that rationality requires us to accept a strict version of *Critical Present-aim Theory* in which morality and rationality coincide.[15]

There are three obstacles on the road towards the unified theory of morality and rationality, and Parfit is determined to overcome them in his work. The prerequisites for the formulation of the new theory are that

1) the commonest concept of rationality (*Self-interest theory S*) is refuted,

[13] This is not just the case with Christianity, but appeals to most Muslims, Buddhists and Hindus as well. Parfit 1984 130. See also Parfit 1984, 194. Somewhat astonishingly, the alliance of morality and self-interest on the basis of the belief in a punishing god and an eternal hell is all that Parfit allows the history of Christian moral philosophy to amount to. For a different view see, e.g., Griffin's (1990, 128–129) remark on the relation between morality, self-interest, and religious belief.

[14] Parfit 1984, 112, 113–114, 194.

> or shown to be no more rational than an alternative view coinciding with the demands of morality;[16]
> 2) we abandon the common-sense non-reductionist theory of personal identity, although nearly everyone naturally assumes it, and adopt the true, reductionist view of personal identity;[17] and that
> 3) we build a new view of morality and rationality on the reductionist concept of personal identity.[18]

In reference to the first point on this programme towards a new understanding of morality, Parfit presents three arguments to refute the self-interest theory *S*. First, *S* is an indirectly individually self-defeating theory; second, there is another theory of rationality *CP*, which is, if not more rational, at least *no less* rational than *S,* but which does not conflict with morality; and finally, *S* is dependent on a mistaken theory of personal identity. In the following, I will first discuss Parfit's formulation of *S*, and then briefly examine his argument for the case *S* against *CP*. I will not discuss Parfit's first arguments against *S*, — that it is indirectly individually self-defeating — for he acknowledges that this argument does not actually refute the theory. Moreover, this topic does not concern our main theme.[19] I concentrate on Parfit's discussion of personal identity, and its implications for moral theory. This involves presenting points two and three on Parfit's "programme". I refer to some of Parfit's fanciful, spectacular examples, so characteristic of his style. Finally, I estimate critically Parfit's ethical theory in the light of the concept of a person

---

[15] Parfit 1984, 443. The way in which Parfit presents *CP* makes it difficult to follow the actual outline of his argument: he lets *CP* take different forms depending on the counter-argument from theory S against criticism from *CP*. Parfit is sure not to have faced a total failure in the theory; he is convinced that a weaker version of his argument survives, according to which it is not less rational to follow *CP* than it is to follow the until now dominant theory of rationality *S*. Parfit 1984, 126. Parfit also claims to have succeeded in establishing some facts which enable him, or some other moral philosopher, to construct the combined theory of morality and rationality in the future. Parfit 1984, 443, 453–454.

[16] Parfit 1984, 112, 113–114, 119.

[17] Parfit 1984, 215.

[18] Parfit 1984, 284–285, 317, 345–347.

which his theory implies. But first we have to examine Parfit's arguments for refuting the self-interest theory *S*.

1.3.2. How the self-interest theory of rationality is to be refuted

According to Parfit, all theories answering the question what we have most reason to do tell us *what we are to try to achieve*. Now, according to moral theories, we ought to try to act morally, whereas according to theories about rationality, we ought to act rationally. Stated in this way these theories give us what Parfit calls our *formal aims*. Different moral theories, and different theories about rationality, again, give us different *substantive aims*. So if we decide to follow a theory about rationality, rationality will be our formal aim, and the specific theory of rationality we choose to guide our action will then spell out our substantive aim. The same applies to moral theories. Parfit examines what in his regard is the commonest theory of rationality — *Self-interest Theory S*.[20] According to him, for all versions of *S* the following hold true:

> (*S1*) For each person, there is one supreme rational ultimate aim: that his life go, for him, as well as possible.[21]
> (*S2*) What each of us has most reason to do is whatever would be best for himself.

---

[19] "Reconsider the Self-interest Theory. This tells each to do the best he can for himself. We are discussing cases where, if we all pursue self-interest, we are doing what is worse for each. The Self-interest Theory is here directly collectively self-defeating. But we cannot assume that this is a fault. [...] *S* is universal, applying to everyone. But *S* is not a collective code. It is a theory about individual rationality. [...] *S* is individually successful. Since it is only collectively self-defeating, *S* does not fail in its own terms. *S* does not condemn itself." Parfit 1984, 92.

[20] Parfit 1984, 3.

[21] Parfit 1984, 4.

> (*S3*) It is irrational for anyone to do what he believes will be worse for himself.
> (*S4*) What it would be rational for anyone to do is what will bring him the greatest *expected* benefit.[22]

If we follow Parfit's terminology, the list provides us with the *substantive* aims of *S*.[23] Furthermore, for explicating what it means to achieve these aims a theory about *self-interest* is needed; this theory will then explicate what is meant by one's life going as well as possible.[24] According to Parfit, there are three versions of self-interest theories, *hedonistic theory*, *desire-fulfilment theory*, and *objective list theory*.[25] One of Parfit's central claims is that (*S1*) – (*S4*) apply to all these three theories. But as his purpose is not to discuss the content of self-interest, he leaves aside the matter of which of the three is best. His argument only concerns the theory of rationality *S*, not the best or most plausible definition of self-interest.[26]

The theory *S*, as Parfit formulates it, is a theory of the *rationality of aims*, and not one concerning the *rationality of means*, as we normally assume when we speak about rationality; this makes it difficult to discern what is really at issue here.[27] Moreover, in discussing the *Self-interest Theory*, Parfit maintains that the substantive rational aim of each version of *S* is that we seek the best outcome for our lives. He does not discuss how we should define the "best for oneself" as our substantive aim, but simply introduces three versions of self-interest as viable interpretations for what "as well as possible" means in a human life. These interpretations provide us with a moral theory, and connected with the self-interest theory, they give us different versions of how we should understand what living a good human life involves. These variants include egoistic as well

---

[22] Parfit 1984, 8.

[23] See Parfit 1984, 3, 45, 48.

[24] Parfit 1984, 3-4.

[25] For Parfit's complete discussion of different versions of these theories, see Parfit 1984, 493–502.

[26] Parfit 1984, 4.

[27] Parfit 1984, 45–49 attempts to deal with the issue, but proves unable to clarify the points discussed here.

as altruistic theories. Now, one of the cornerstones of Parfit's argument against *S* is that it conflicts with morality, and this conflict gives rise to the attempt to combine morality and rationality within one unified theory. However Parfit's claim about the necessarily egoistic nature of the Self-interest theory is false even in terms of his own definition; altruistic versions of *S* are not incompatible with morality.[28] Moreover, the weight of his argument against *S* in the second part of his book rests on the assumption that all versions of *S* disregard the good of others for the sake of the agent's own good. This claim is not only false but it is also a new interpretation of *S*, diverging from Parfit's original (*S1*) – (*S4*) theses.[29] These obscurities and muddles weaken his reasoning considerably; it is difficult even to discern what argument he is trying to defend, and what his point of criticism actually is.

There might, however, be an interpretation which can make Parfit's ideas more intelligible. His main aim seems to be to change the perspective from which we examine our substantive aims, that is to say, the horizon against which we regard ourselves as moral agents. Parfit is a utilitarianist and his utilitarian definition of the good, namely that satisfying one's preferences makes one's life go well for oneself, seems to contradict morality. Parfit does not consider reformulating his utilitarian aim. Instead he suggests that we must change the perspective from which we see this goal. This means that the crucial point of Parfit's criticism against *S* is that *S* must be abandoned because it focuses attention on *one's life*. According to this interpretation, Parfit tries to design an alternative approach to moral questions, or a new framework within which we are to review our actions, and within which morality is to be applied. To estimate if this is a tenable interpretation of Parfit's aims, and

---

[28] Parfit seems to acknowledge this (see Parfit 1984, 466–467), but this happens in a context where he examines *S* in relation to a rival theory of rationality, the present aim theory *P*, not in regard to morality.

[29] The initial form of *S* is "For each person, there is one supremely rational ultimate aim: that his life go, for him, as well as possible." (Parfit 1984, 4), and the version of *S* Parfit claims to have defeated in the second part of his book "insists that a rational agent give supreme weight to his own self-interest, *whatever* the costs to others. It insists that a rational agent must be biased in his own favour." Parfit 1984, 192.

what the alternative framework for morality might be, we must examine his argument further. Let us leave this question until later, and now move on to scrutinize the "duel" between *S* and an alternative theory of rationality, *CP*.[30]

Parfit's suggestion for the theory of rationality that is to replace *S* is *CP*, *Critical Present-aim Theory*. It maintains that

> Some desires are intrinsically irrational. And a set of desires may be irrational even if the desires in this set are not irrational. [...] A set of desires may also be irrational because it fails to contain desires that are rationally required. Suppose that I know the facts and am thinking clearly. If my set of desires is not irrational, what I have most reason to do is what would best fulfil those of my present desires that are not irrational. This claim applies to anyone at any time.[31]

We can, however, define two stricter derivative versions of *CP*, which, if provided with reasons strong enough, make rationality coincide with morality:

> (*CP1*): Each of us is rationally required both to care about morality, and to care about the needs of others. Since this is so, we have a reason to act morally, even if we have no desire to do so. Whether we have a reason to act in a certain way usually depends on whether we have certain desires. But this is not so in the case of desires that are rationally required.[32]
> (*CP2*): There is at least one desire that is not irrational, and is no less rational than the bias in one's own favour. This is a desire to do what is

[30] I have taken the liberty of reducing Parfit's discussion on the two theories to one single argument for the purpose of introducing the main point but skipping the muddles. Consequently, I leave out the discussion on temporal neutrality vs. temporal relativity and the attack against *S* on formal grounds. The exposition of Parfit's line of thought does not get impaired, however, because Parfit's whole discussion is based upon a reformulation of *S* which deviates from his initial position.

[31] Parfit 1984, 119. Parfit maintains that the theory *CP* he advocates is actually Richard Brandt's theory of rationality concerning the rationality of desires and actions. Parfit 1984, 118–119, 512, n. 2.

[32] Parfit 1984, 121–122.

> in the interests of other people, when this is either morally admirable, or one's moral duty.[33]

Parfit's formulations include an implicit conception of both intentional action in general and of moral action. Parfit represents, with Brandt and Hare, the view that desire is the moving force of action. He follows Brandt in thinking that people can estimate the rationality of their desires, although he does not develop this thought further. We can differentiate between three kinds of intentional action. First, there is intentional action as such, which takes the form of practical syllogism. Second, there is intentional action which is based on rational desires. And the third category of intentional action is moral action, which is a subcategory of the second in the sense that it is based on rational desires. Moreover, intentional action becomes morally relevant when it concerns the needs and interests of other people.

Parfit's main argument in defence of *CP* and against *S* seems to depend on the possibility of finding cases in which (*CP2*) is true.[34] Such cases are crucial for Parfit's argument against *S*, at least if we accept that for all versions of *S* it is always irrational not to be biased in one's own favour.[35] *CP* includes an implicit claim that altruism is a natural human quality. Given its naturalness, *CP* holds it rational to be altruistic.[36] Parfit uses this assumption against *S*: when we ask *S* whether it is rational to adopt an altruistic attitude, we do not get a reply. *S* cannot justify the rationality of altruism, because it is a second-order question about rationality, and answering this would involve using *S* for justifying *S* itself. *S* cannot be used for giving grounds to the rationality of choosing

---

[33] Parfit 1984, 131.

[34] Parfit presents such a case: "*My Heroic Death*. I choose to die in a way that I know will be painful, but will save the lives of several other people. I am doing what, knowing the facts and thinking clearly, I most want to do, and what best fulfils my present desires. [...] I also know that I am doing what will be worse for me. If I did not sacrifice my life, to save these other people, I would not be haunted by remorse. The rest of my life would be well worth living." Parfit 1984, 132.

[35] Parfit 1984, 192.

[36] In this point Parfit's view accords with that of Brandt's see, page 42.

S, instead of some other theory. Consequently, it is not *less rational* to choose some other theory to outline the criterion for rationality. In case we might not be convinced by Parfit's argument, he presents a very different way of refuting *S*; so let us now move on to that.[37]

1.3.3. The reductionist view of personal identity

The concept of personal identity is needed to explain what the nature of a person is: what it is that makes somebody one and the same person at two different times; to explain the continuity of the stream of consciousness, and to illustrate how the life of a person can be experienced as a unity.[38] The discussion Parfit presents to clarify and to reformulate the concept of personal identity is, unfortunately, confused. This is mainly due to the fact that he uses the central term 'personal identity' in two ways. First, he signifies a person's being the same person at two different moments of time by calling that "our identity over time". This use of the concept implies that there is an unambiguous meaning of "personal identity", which can be used to explain that a person remains the same over time, or that she continues to be herself, and what that entails.[39]

---

[37] See the section 'Personal identity and rationality', Parfit 1984, 306–320.

[38] Parfit 1984, 202, 214.

[39] "They are Reductionist because they claim (1) that the fact of a person's identity over time just consists in the holding of certain more particular facts, and (2) that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists." Parfit 1984 210; and: "Our identity over time just involves (*a*) Relation R — psychological connectedness and/or psychological continuity [...] Personal identity is not what matters. What fundamentally matters is Relation R, with any cause." Parfit 1984, 216, 217. And still: "Since these views [Reductionism and Non-Reductionism] disagree about the nature of persons, they also disagree about the nature of personal identity over time. On the Reductionist View, personal identity just involves physical and psychological continuity." Parfit 1984, 275. Parfit has also called the third part of his book 'Personal identity', as if he were giving a new, unarbitrary formulation to this concept.

Second, he seems to repudiate the term completely, as one applying only to the non-reductionist view of personal identity, a view he attempts to disprove: "Personal identity is not what matters".[40] It is not always clear which of the two Parfit means when he develops his argument. It seems that he tries to show that the suggested concepts of personal identity do not square with facts about the world, and break with the best philosophical models for understanding the nature of ourselves.[41] He seems to claim that because personal identity is not what matters, there must be something else that matters instead; in order to answer the crucial question whether we will, in some puzzle cases, die or not, we do not need the concept of personal identity. Our search for a concept of personal identity already shows that we have a totally mistaken view of ourselves and of our nature as persons.[42]

Parfit maintains that the commonest view of personal identity is *non-reductionism*. Non-reductionism involves accepting the following:

1) persons are separately existing entities;
2) personal identity is crucial for the continued existence of a person and psychological connectedness and continuity must be explained with the help of one subject having all the experiences that constitute the psychological continuity and connectedness; and
3) personal identity is always a determinate further fact.[43]

Psychological connectedness means "the holding of particular direct psychological connections", and psychological continuity indicates "the holding of overlapping chains of strong connectedness".[44]

A cornerstone of Parfit's argument is that all these three views are

[40] Parfit 1984, 215, 217. In adopting this usage of the word, to entertain the idea of personal identity might even be nonsensical; there seems to be no concept of personal identity meeting all the requirements for it. See the discussion in section 91; Parfit 1984, 266–272.

[41] This claim relates to the discussion in Williams 1973, 19–25. For full details, see Parfit 1984, 267–273.

[42] Parfit 1984, 216. See also section 94. 'Is the true view believable?', Parfit 1984, 274–280.

[43] Parfit 1984, 210, 216.

insolubly bound together: we cannot hold one of these claims, unless we are also ready to accept the other two.[45] The strategy, then, is to prove, with some examples, that the first point is false, that persons are not separately existing entities. If this is true, points 2) and 3) lose their weight, too.[46] This approach makes the analysis susceptible, though. Not all non-reductionists are ready to accept that persons are separately existing entities, although they hold out for the truth of point 2). Parfit's analysis only covers a small part of the non-reductionist argument.[47]

The central thesis of the argument is that most of us adopt the three non-reductionist premises as the truth about ourselves and others, and this explains partly why we think it rational to act according to our self-interest: because personal identity is seen as an always determinate "deep further fact"; it makes a difference whether or not we act in accordance with our self-interest. From our own point of view, this is perhaps the only thing that really matters.[48] Although understandable, the non-reductionist view of personal identity is false.[49]Additionally, it is a great hindrance for accepting a new understanding of ethics, a view that helps to combine the projects of rationality and morality. Instead, we should adopt the true view — reductionism. If we accept reductionism, we also have to forsake other beliefs combinable only with non-reductionism.[50]

Parfit presents two variants of non-reductionism, one which he calls

---

[44] Parfit 1984, 206. Psychological connections are said to involve direct memories, the connections holding between an intention and the later act in which the intention is carried out, and connections which hold when a belief, or a desire, or any other psychological feature, continues to be had. We may speak about psychological continuity, ie. strong connectedness, when there are enough direct connections. Parfit does not define what would count for "enough". Parfit 1984, 205–206.

[45] Parfit 1984, 216

[46] Parfit 1984, 217. Ricœur (1990, 157–158) points out that Parfit's definition of non-reductionism is actually parasitic of reductionism, given in the Davidsonian vocabulary of facts and events; see Donald Davidson *Essays on Actions and Events*. Clarendon Press, Oxford, 1980.

[47] Geoffrey Madell (*The Identity of the Self*. Edinburgh University Press, Edinburgh, 1981) represents this kind of a view. According to him, we must think of ourselves as being irreducible subjects, but this does not mean that we would have to postulate a separately existing ego; see Madell 1981, 134–139.

[48] Parfit 1984, 279–280.

radical or *Cartesian*, the other *moderate non-reductionism*. The Cartesian alternative presupposes that there exists a separate entity, "self", and that the existence of this self is an all-or-nothing fact. The moderate variant of non-reductionism states that personal identity is not a separately existing further fact, but that it is still a non-additive characteristic, something more and further than just the psychological and physical facts about a person. Both these views hold that not only psychological continuity but also unity within a life must be explained in terms of personal identity. Personal identity determined by the non-reducible human self is the only *explanans* needed whenever we want to find an explanation for the continuity and unity of an individual's life and experiences.[51]

To show that what we normally think is false, Parfit takes up some extraordinary examples. It is not possible for us to test these cases chosen in support of his philosophical views but, Parfit claims, this is merely a technical, not a logical impossibility.[52] The legitimacy of these examples

---

[49] Parfit finds non-reductionism a natural standpoint: this belief is caused by evolution. During some indefinable period of the existence of the human species this conviction served the survival of human beings. Although understandable in this light, our intuitions about the nature of our identity lead us astray. Parfit 1984, 279–280, 308. For non-reductionism to be true there should be scientific, and philosophically credible evidence for it, e.g., presupposing also reincarnation, we could remember incidents from our former lives, and the veracity of these memories could be tested. Parfit 1984, 227. Or we could state, in the case of brain operations, that an operation would either change or destroy the whole personality of a person, or leave it intact; and that there would be no other alternatives. We have, however, no evidence of either kind, and therefore, we should abandon non-reductionism as a false conviction. Parfit 1984, 227–228. It is not quite clear what Parfit actually aims at with this example. He speaks about the non-reductionists holding a view according to which a self is a separately existing, spiritual substance. So, what evidence could a brain operation, whatever its effects, produce against the existence of such a substance?

[50] "If we cease to believe that our identity is what matters, this may affect some of our emotions, such as our attitude to ageing and to death. And, as I shall argue, we may be led to change our views about both rationality and morality." Parfit 1984, 215. See also Parfit 1984, 347.

[51] Parfit 1984, 210, 214. Parfit does not mention any philosophical "commissioners" for his categories of radical and moderate non-reductionism. Richard Swinburne could be seen as a representative of the radical branch, whereas Geoffrey Madell could be classified as defending a moderate form of non-reductionism; see Swinburne's account in Shoemaker & Swinburne 1984, and Madell's in Madell 1981. About classifications into these categories, see Sprigge 1988.

or "thought experiments" lies in the fact that they compel us to abandon our common ways of thinking. We cannot follow the usual paths of our intuition, but have to search for the best philosophical alternative to answer these puzzling dilemmas.[53] We can, however, invalidate this justification for the use of imaginary arguments as well as undermine the claim that the constraints of testing these examples are merely technical. Adopting Parfit's assumptions means that we in fact accept the conclusion as true before anything has been proved. In other words, we cannot make sense of Parfit's examples, unless we adopt his version of personal identity as a description of what takes place in his test cases. As Parfit's critics, we should, however, not only claim that his puzzles are technically inviable, but that they are also philosopically defective in the sense that they incorporate the models they are arguing for as implicit presuppositions.[54]

---

[52] Parfit 1984, 200, 234, 255.

[53] Parfit defends his use of imaginary cases as examples for supporting philosophical views: "This criticism [against the use of such examples] might be justified if, when considering such imagined cases, we had no reactions. But these cases [that Parfit uses himself] arouse in most of us strong beliefs. And these are beliefs, not about our words, but about ourselves. By considering these cases, we discover what we believe to be involved in our own continued existence, or what it is that makes us now and ourselves next year the same people. We discover our beliefs about the nature of personal identity over time. Though our beliefs are revealed most clearly when we consider imaginary cases, these beliefs also cover actual cases, and our own lives." Parfit 1984, 200.

[54] Parfit considers the possibility of his examples being, not only technically, but also "deeply impossible", but he passes over this alternative as irrelevant for his argument: Einstein's thought-experiment of asking what he would see if he could travel beside some beam of light at the speed of light, is deeply impossible, but in no ways non-sensical. Parfit 1984, 219. Parfit's defence for his use of imaginary cases misses the point, however. He does not even consider the possibility that the very formulation of his phantasy examples already includes deep-rooted presuppositions which concern the topic at issue, the concept of personal identity.

Let us consider *Teletransportation*. Teletransportation is a way of travelling with the speed of light. The teletransporter first records the exact states of the traveller's brain and body, destroys them, and then transmits the recorded information by radio to the traveller's destiny where a replicator will create, out of new matter, an exactly similar brain and body as the traveller had on Earth. The traveller who has fallen unconscious during the procedure now wakes up.[55] Who is the person waking up in this body, Parfit asks; the one who entered the teletransporter or someone else, and if someone else, then, who? Parfit's answer is: for every non-reductionist theory, personal identity teletransportation and suchlike phenomena remain an unsolved puzzle.[56]

It is difficult to see what non-reductionists might actually have against Parfit's description of teletransportation. To take an example, one version of non-reductionism states that psychical phenomena are combined together with the brain and its functions, but that they are not, in any way known to us, reducible to purely physical manifestations. According to this view, teletransportation would just transport a person's "soul" to Mars. Such views are not unknown to Parfit, but he leaves them out of this context, because the arguments supporting those views "are at a very abstract level".[57] Although Parfit insists that only reductionists are able to explain what happens in teletransportation he does not explicate

---

[55] Parfit 1984, 199–200.

[56] Parfit 1984, 242, 243.

[57] "[...] some writers reject *both* this last Reductionist claim *and* the Cartesian View. These writers do not believe in Cartesian Pure Egos. And they do not believe that a person is any other kind of separately existing entity. They believe that the existence of a person just consists in the existence of his brain and body, and the doing of his deeds, and the occurrence of various other physical and mental events. But these writers claim that we cannot refer to particular experiences, or describe the connections between them unless we refer to the person who has these experiences. On their view, the unity of a mental life cannot be explained in an impersonal way. Strawson discusses an argument for this view, suggested by Kant. This argument claims that we could not have knowledge of the world about us unless we believe ourselves to be persons, with an awareness of our identity over time. Shoemaker advances a similar argument. If these arguments are correct, they might refute my claim that we could redescribe our lives in an impersonal way. Because these arguments are at a very abstract level, I shall hope to discuss them elsewhere." Parfit 1984, 225.

what the philosophical point behind this test case is; that changing the matter of a human body does not affect a person's psychological qualities.

Reductionism denies that the subject of experience is a separately existing entity, distinct from a brain and body, and a series of physical and mental events. Therefore, a person's existence only consists of a brain and a body, and in the occurrence of a series of interrelated physical and mental events. Everything that we say about persons could be reduced to language including no concept of a person, but only conceptions describing physical events in the brains and bodies of human beings.[58]

If we accept Parfit's view, we can characterize human intentionality by saying that there are two approaches to these phenomena: first, the physical processes which take place in human (brain)cells form a "micro level", and second, the mental events constitute a "macro level" to the phenomena. We can refer to events on the macro level by regarding human beings as intentional agents. This is only a way of speaking, though. All occurrences on the macro level are reducible to the events on the physical micro level. The language of intentionality is not enough for describing a person, whereas a description of phenomena on the micro level is sufficient for a total understanding of what takes place in a human being. The difference between Parfit's reductionism and non-reductionism is that the latter considers human intentionality as a non-additive characteristic which cannot be reduced to the material without something essential being lost, whereas Parfit sees that the reductionist explanation of mental phenomena captures everything essential for understanding them.[59]

What is the right explanation for teletransportation? Parfit's answer is

---

[58] Parfit 1984, 211, 212–213, 223.

[59] Shoemaker who counts himself as one of the reductionists, and who gives Parfit great credit for his work, criticizes him for going too far: "I suspect that if reduction requires an 'impersonal description' of the sort Parfit sometimes seems to have in mind, it will be impossible to have a reduction of personhood and personal identity without having a reduction of mentality as well — the impersonal description will have to be in physical or functional terms." Shoemaker 1985, 446–447.

that it is not personal identity based on a further fact, but only *Relation R*, by which he means *psychological connectedness and/or continuity* with any kind of cause. In the case of teletransportation, we cannot pin down anything on the basis of which we could say that this or that thing either would or would not preserve our identity. This means that the whole question concerning personal identity is misguided. It is basically an unintelligible problem for we are not separately existing entities apart from our brains and bodies. Besides, it does not even matter for us as persons whether our identity remains intact. The real question that interests us is whether our existence continues or not. All we really need to know in cases like teletransportation is whether *Relation R* holds or not, that is, whether there is psychological connectedness and/or continuity between what we now are and what we were at some earlier moment in time, or between what we now are and what we will be at some later moment in time. In the case of teletransportation, Relation R holds so we can count on it for survival.[60]

Parfit offers more examples to back up his theory. The next thought-experiment, called *Combined Spectrum*, the combination of physical and psychological changes taking place in a person simultaneously, is supposed to prove that

> we cannot defensibly believe that our identity involves a further fact, unless we also believe that we are separately existing entities, distinct from our brains and bodies. And we cannot defensibly believe that our identity must be determinate, unless we believe that the existence of these separate entities must be all-or-nothing.[61]

The two unities — the unity of consciousness at any time, and the unity of a whole life — are not explained by saying that unity is constituted by the fact that the same person undergoes different experiences. These unities must be explained by describing the relations between these many

---

[60] Parfit 1984, 271, 275, 279–280.

[61] Parfit 1984, 240. Here again Parfit speaks about personal identity as if it were not a problematic concept for him.

experiences, and their relations to this person's brain. This can be done with the help of *Relation R*, compatible with reductionism.[62]

> *Combined Spectrum*: A group of scientists has recorded all the states of all the cells in Greta Garbo's brain and body, as they were when she was 30. In this experiment all the cells of the body of Derek Parfit are being replaced by new organic matter, by cells the genetic and other information of which accord with that of the cells of the 30-year-old Greta Garbo. This experiment is being accomplished slowly, with the speed of normal cellular renewal that takes place in every human being. At the beginning of the test the person will be fully continuous with Derek Parfit, both physically and psychologically. Then, a few of the cells in his brain and body will be replaced. This will not cause a notable difference. Only gradually the memories, traits of character and physical appearance of Derek Parfit change. He will have some apparent memories of Greta Garbo's life, and have some of Garbo's characteristics, as well as some of her physical features. As the experiment is carried through, the resulting person will resemble Derek Parfit less and less, but will, to an increasing degree bear likeness to Greta Garbo. At the end of the experiment the resulting person will be identical with the 30-year-old Greta Garbo.[63]

According to Parfit's interpretation, if we are supporters of a non-reductionist theory of personal identity, we cannot describe the experiment as it is illustrated here. As non-reductionists, we are doomed to say that there is either some critical percentage *x* of cells before the replacement of which no change can be noticed in Derek Parfit; or that there is some part of the brain which the "self" inhabits, and the removal of which is the only thing that affects the personality of the person in question.[64]

---

[62] "Reductionists claim that nothing more is involved in the unity of consciousness at a single time. Since there can be one state of awareness of several experiences, we need not explain this unity by ascribing these experiences to the same person, or subject of experiences. [...] In explaining the unity of this life, we need not claim that it is the life of a particular person. We could describe what, at different times, was thought and felt and observed and done, and how these various events were interrelated." Parfit 1984, 251. See also Parfit 1984, 263.

[63] Parfit 1984, 236–237.

We cannot make this experiment, but Parfit is convinced that his description would be right if the experiment were possible. This reveals an important, implicit prerequisite underlying Parfit's reductionism, namely the conviction that mental phenomena are bound to individual cells. Consequently, physical changes occurring at the micro level of human cells can be directly traced as specific psychological and mental events of the macro level: substituting some of Derek Parfit's brain cells for those of Greta Garbo's would destroy some individual memory of Parfit's life, replacing it with a detail from Greta Garbo's memory.[65] This description presupposes, however, that memories, traits of character, skills, and so forth, exist separately and independently of each other, not only in a person's brain, but also as parts of her mental life. It seems that being Derek Parfit only involves being able to list up different experiences that have occurred to the human being called Derek Parfit during the time of his existence, and that this listing constitutes psychological continuity. Contrary to Parfit's claim, neurophysiological experiments do not support his view. It is neither possible to locate certain higher functions of the brain at some definable point nor to say how these functions are carried through.[66]

There are other problems embedded in the Combined Spectrum. What, in this experiment, establishes the unity of consciousness, something that must be explained with the *Relation R*?[67] And how should the

---

[64] "We might continue to believe that our identity must be determinate. [...] We would then be forced to accept the following claims: Somewhere in this Spectrum, there is a sharp borderline. There must be some critical set of the cells replaced, and some critical degree of psychological change, which would make all the difference. If the surgeons replace slightly fewer than these cells, and produce one fewer psychological change, it will be me who wakes up. If they replace the few extra cells, and produce one more psychological change, I shall cease to exist, and the person waking up would be someone else. There must be such a pair of cases somewhere in this Spectrum, *even though there could never be any evidence where these cases are*." Parfit 1984, 238–239. See also Parfit 1984, 235.

[65] In justifying the use of Combined Spectrum as an example, Parfit writes: "Since our psychological features depend on the states of our brains, these imagined cases [of Combined Spectrum] are only technically impossible. If we could carry out these operations, the results would be what I have described." Parfit 1984, 238.

[66] See the reports of various experiments and conclusions drawn from them in Churchland 1989, 217–235.

psychological continuity and the unity of a person during this lengthy experience be conceived? Would he/she have an understanding of him/herself as someone? The person would also have experiences, feelings, etc., during the experiment, all of which will — to use Parfit's way of describing the case — leave a mark on the brain cells, part of which will be the transplanted cells of Greta Garbo. The resulting person would not be Greta Garbo at the age of 30, but a person remembering, perhaps somewhat vaguely, not only some of the incidents of Greta Garbo's life, but also the experience of going through the experiment. Accordingly, it would not be out of the question for the person to say, when the whole experiment was completed, that she is Derek Parfit. She would not really remember anything from the life of the former, original Derek Parfit, but it would make sense to say that she was Derek Parfit, whose sex, appearance, memories, and so forth, were all changed to those of Greta Garbo's. This would not, however, make her Greta Garbo. Psychological unity does not consist of the ability to list experiences linked together chronologically. What *Relation R* cannot provide is an over-all view of oneself and of one's experiences amounting to a life-story, something that many philosophers find crucial for our understanding of ourselves as persons.[68]

There is an amazing feature in Parfit's theory which the Combined Spectrum example displays clearly. Parfit is a reductionist who claims

---

[67] Oaklander (1987, 527–529) claims that Parfit's criteria for a continued existence of a person do not account for the unity of a person's life at all.

[68] For the type of non-reductionism, which Parfit passes without further notice, this point of view is of particular interest and importance. Parfit 1984 225. Taylor 1989, 49–50 criticizes Parfit's *Relation R* heavily: "[...]this whole conception suffers from a fatal flaw. Personal identity is the identity of the self, and the self is understood as an object to be known. [...] But what has been left out is precisely the *mattering*. The self is defined in neutral terms, outside of any essential framework of questions. [...] But if my position here is right, then we can't think of human persons, of selves in the sense that we are selves, in this light at all. They are not neutral, punctual objects; they exist only in a certain space of questions, through certain constitutive concerns. The questions of concerns touch on the nature of the good that I orient myself by and on the way I am placed in relation to it. But then what counts as a unit will be defined by the scope of the concern, characteristically, the shape of my life *as a whole*. It is not something up for arbitrary determination." Wolf (1986, 247) directs criticism of a similar kind towards Parfit.

that all mental processes can be reduced to physical processes and all that matters for our survival is the *Relation R*. Despite the materialist tenor of the theory, Parfit does not actually give any real weight to human corporeality.[69] All his examples presuppose that we may imagine parts of a human body, or brain to be freely tampered with, changed, replaced, etc., regardless of whether this is physically possible or not, as long as it is not logically contradictory. Any physical change is accepted as possible if it accords with the view of *Relation R* as the foundation of personal identity. To take an example, what does it really mean to replace Derek Parfit's cells with new organic matter, the genetic and other information of which accord with that of the cells of the 30-year-old Greta Garbo? Parfit describes the case as simply recording the states of all the cells in Greta Garbo's brain and body as they where when she was 30, and then constructing new organic matter on that basis. But is this intelligible? What does such recording mean? What does it mean to construct new organic matter? Do we do this separately, cell by cell, or atom by atom? Or do we first construct a complete "Greta Garbo" which we then, little by little, deconstruct as we use the cells of this piece of organic matter for replacing Derek Parfit's cells? It seems that Parfit has designed his example using the theory he wishes to prove true as a presupposition for what is regarded as possible, or impossible. Seen from this angel, Parfit's examples resemble the traditional Cartesian test-cases, allowing all kinds of destruction, or changes in the body, which then have no effect on the spiritual essence constituting the person. Similarly, Parfit's examples permit any change as long as this does not violate *Relation R*.

Parfit does not stop to consider such difficulties, because for him the Combined Spectrum example is one more piece of evidence against non-reductionism and in favour of reductionism. He claims that we can defend the view that our identity is a separately existing entity only if we simultaneously hold that the self is a "creature" existing independently of our brains and bodies. We must also abandon the claim that the exist-

[69] See Ricœur 1990, 158–159 for a resembling remark.

ence of a separate self is an all-or-nothing fact. Even though we knew exactly how many cells of Derek Parfit had been replaced by those of Greta Garbo's, we would still not be able to answer the question, when that person ceases to be Derek Parfit and begins to be Greta Garbo. The answer depends on the description we choose as accurate for the procedure. Personal identity is not what matters, only *Relation R.*[70]

Parfit introduces still another example to convince his readers of the non-importance of the non-reductionist personal identity and of the all-prevailing meaning of *Relation R;* The Identical Triplet Brothers:[71]

> In cases of severe epilepsy the cutting of the neural cord which combines the two hemispheres has sometimes been used as a cure. After this has been done, it has for some time been the case that the people who have undergone the operation have had two separate streams of consciousness, one in each hemisphere. Normally one of the two takes over after some time, and the effect of two consciousnesses vanishes. Usually it is also the case that the hemispheres are specialized in performing certain functions, but it is not impossible that there exist someone whose two hemispheres both contain the same information and control the same functions.
>
> Let us suppose, for the sake of the argument, that there are identical triplet brothers, *x*, *y* and *z*, who all have brains with hemispheres each containing the same information as the other. Let us also suppose that it is possible to make brain transplants, not only transplanting the whole of a brain successfully, but also a brain split in two. Now, the identical triplets are all injured lethally. The injury has affected *y*'s and *z*'s brains only, which are rapidly being destroyed, while their bodies have survived without any damage. *x*, on the other hand, has been exposed to a different kind of misfortune: his brain is intact, but due to the injury his body has undergone, he will soon encounter a premature death. Luckily enough, all three are quickly brought into a hospital where the doctors bisect the

[70] Parfit 184, 240.

[71] The following section is a summary of Parfit's discussion and examples from Parfit 1984, 245–246, 250–251, 253–255. The example Parfit uses (in a modified form) derives from Williams 1973 and Wiggins 1967; see Puccetti 1980, 583; Shoemaker 1985, 444; Parfit 1984, 254, 518.

> undamaged brain of *x* and transplant each resulting half into the unimpaired skulls of *y* and *z*. So two people survive.

"Who are the two people surviving this accident?" Parfit asks. They can possibly be neither *y*, nor *z*, because they are both psychologically wholly continuous with *x*. But to say that they both are *x*, Parfit maintains, would violate too strongly against our concept of a person. Astonishingly, Parfit now finds it unproblematic to state that this interpretation "distorts the concept of a person", as if there were some unarbitrary, common understanding of the concept and of its bearings, forgetting that redefining that very concept is the most central issue of his discussion.[72] According to Parfit, the remaining persons would both be psychologically (and to a great extend physically) similar to the former brother *x*, so neither would have more right than the other to claim to be the "real" *x*. The best solution, then, would be to regard neither of the surviving people as *x*. But, if *x* had time before the operation to consider whether to accept the offer of the doctors or not, Parfit is sure that *x* should view his future prospects as two different people being at least as good as a normal survival as one person. The relation of *x* to each of the surviving individuals contains everything that can sensibly be required from continued existence. The only problem is that there will be two future people continuous with *x*. But this would not mean death for *x*, on the contrary, the prospects of such continued existence would surpass those of conventional survival.[73]

The interpretation Parfit gives to this thought-experiment seems problematic. It presupposes that the past and future experiences of a person are related to her present self in the same way; so it is supposed that *x* is able to consider his survival as two people in the future in the same way as he would think about his past. We could object to Parfit's description of *Combined Spectrum* in that thinking about the past involves

---

[72] "Could we still claim that I survive as both? [...] I might say: 'I survive the operation as two different people. They can be different people, and yet be me, in the way in which the Pope's three crowns together form one crown.' This claim is also coherent. But it again greatly distorts the concept of a person. [...] it is hard to think of two people as, together, being a third person." Parfit 1984, 257.

not just remembering separate incidents of the past but an understanding of one's present self through everything one has experienced and lived through. Thinking about the future is different; even if one could know that certain things will most likely happen to one, one could still not necessarily see how that would effect one's interpretation of life. In the case of the triplet brothers, the two surviving people would both claim to be *x*. According to Parfit, we must abandon this solution at once.[74] But this interpretation will only fit Parfit's scheme if we think that the relation between a person's past experiences and her present self on the one hand, and the relation between her future experiences and her present self on the other hand, are identical.[75] For the undivided *x*, the prospect of surviving as two is, naturally, impossible to conceive, but so would be the thought of being paralyzed, say, for a world record sprinter. The two surviving people would both think of *x*'s past as their own past, and their understanding of themselves would be that of the undivided *x*. They could both be said to have been the same person as the undivided *x*, but it would still not be right to say that the undivided *x* were the same human being as either or both of them.

One of the problems with Parfit's discussion is that he does not dis-

---

[73] Parfit 1984, 261, 263–264. Gillet (1986, 76–78; 1987, 224–229) criticizes Parfit for his interpretation of what happens in brain bisection. Parfit adopts the hypothesis about the effects of brain bisection from Sperry 1966 without questioning them at all. Speaking about a person having two separate consciousnesses at a time is misconceived, though. "In real life the state of informational disruption [after the severing of most of the fibres between the two hemispheres] is a transient phenomenon and the patients tend to 'get their act back together again' albeit gradually. If such people do function in such a remarkably integrated way, surely we should ask 'Who is performing these tasks?', or, 'Is there a single subject of experience who is making and is aware of making these mistakes?' rather than 'How many separate streams of consciousness are there here?' The fact is that the person realises that *he* has made a mistake, not that someone, perhaps contingently related to him, has made a mistake which he has the knowledge to correct. As far as he is concerned, the person in error and the person who is not are he, himself, one person and one mind, but he is not functioning properly." Gillet 1986, 226. According to Gillet, the whole use of Parfit's thought experiments is deeply susceptible. Gillet 1986, 229.

[74] Parfit 1984, 258.

[75] This is even clearer in part three of his book where he discusses past and future suffering. Parfit 1984, 165–168, 181–184.

tinguish between being a person, being an individual, and being a human being. Neither does he discuss the difference between "being the same person as someone", and "being personally identical with someone". He sticks to the logic of identity, and finds our common-sense views nonsensical. If he spoke about present persons being the same as some past persons instead of keeping to identity language, he might get further. There is also another feature in the terminology which complicates the discussion. After his declaration of non-subjective reductionism, Parfit moves from the third-person perspective to a first-person view. He considers all his examples from the point of view of what they would involve personally for him as Derek Parfit. Consequently, "the important question is not, 'Which is the best description?' The important question is: 'What ought to matter to me?'"[76] The first-person perspective, and the construction of the examples from a reductionist point of view are incompatible with each other, and this connection of the two makes our ordinary assumptions about personal identity seem absurd. Although we have been given all the facts of a case we cannot find an answer to our most burning questions.[77]

What, then, does matter for us? What is it that essentially makes us the persons we are? The "substance" of human personhood is given in *Relation R*; psychological continuity and/or connectedness. A person is fundamentally a set of co-existing desires, and this is the only thing that matters in any imagined puzzle-case. Questions which fall out of the scope of *Relation R* do not matter, because we are not that kind of entities these questions presuppose us to be.[78] If the conditions of *Relation R* prevail we know everything there is to know. We do not need anything

[76] Parfit 1984, 260. For Ricœur (1990, 163–166) this feature in Parfit's terminology reveals that the perspective of selfhood, what something is for me, cannot be ignored if we are to give a full account of the concept 'person'. Parfit's reductionism would actually require an impersonal, third-person description of what happens in the puzzling cases, but he constantly uses the first-person mode instead, weighing whether something would matter, or be for "me".

[77] This is how Ricœur formulates the difficulty Parfit's cases exemplify, see Ricœur 1990, 162–163.

[78] Parfit 1984, 264.

apart from these conditions for determining whether something counts as survival or not. This is what matters.

1.3.4. The meaning of reductionism to morality

The example of the identical triplet brothers shows irrevocably, Parfit maintains, that personal identity is not what matters but only *Relation R*. It shows why, from *x*'s point of view, it does not have any meaning, or has only very little meaning, that neither of the surviving people can be said to be *x*. It is a further piece of evidence in favour for reductionism introduced by Parfit: we are not separately existing "selves" for whose lives personal identity is an all-or-nothing fact. The only things relevant for determining the dilemmas of someone's continued existence are established by *Relation R*. Because of this, we must adopt a new view of life: we should change our attitude towards our own death and suffering. Death only means that certain experiences are not linked together in the same way as they used to be. Suffering simply means that certain unpleasant experiences are combined together with certain other experiences. Parfit finds these thoughts comforting.[79]

Because there is no determinate personal identity, psychological continuity and unity within a single life are more weakly established facts than they would be if non-reductionism were true. The weak connectedness combining our different experiences together enables us to regard our own future as a succeeding series of incidents from different persons' futures.[80] Parfit states that this significantly alters the way we understand rationality and morality. Because we actually have no determinate identity, it is not rational to be more concerned about ourselves than about others.[81]

---

[79] Parfit 1984 278–280, 282–285.

[80] Parfit 1984, 313–314.

The truth of *Relation R* also shows that it is irrational to adopt the Self-interest theory *S* for helping us to act rationally. By adopting the reductionist view of ourselves as persons we can combine rationality and morality. Instead of thinking solely about our self-interest and caring just for our own future, we must see life from a reductionist perspective. This means that we must pay attention to our future self — or selves — to the same extent that we pay attention to the prospects of other people. Because identity is not the crucial factor, non-identity is not central either; we must think of other people with the same intensity of affection as we think of ourselves. We may not attend to the question who the subject of some experience is, any more than we may take notice of the time of occurrence of some experience; the questions concerning the subject and time of experiences are both futile in ethics, where the only matter of moral significance is the quality of the experience.[82] Parfit's discussion of personal identity shows that the suggestion made earlier was correct:[83] Parfit's central endeavour in refuting *S* is to provide morality with a new kind of framework. Instead of seeing morality taking form inside *one's whole life*, we should, according to Parfit, consider morality in relation to punctual, contemporarily existing persons, without attaching either a person-constituting past or a future to them.

As we now have explicated Parfit's arguments for refuting what according to him lies behind *S,* and for adopting *CP* we can examine his moral theory and the concept of a person in more detail. Two issues are of special importance here. First, what is the connection between the new perspective *Relation R* is said to provide us with, and Parfit's normative ethical theory? Second, what kind of a metaethical status does *Relation R* have? I claim that Parfit's new perspective on moral questions, to regard "punctual" persons instead of one's whole life, accords with the utilitarian character of his moral theory *CP*. In Parfit's version of utilitarianism the moral aim is the maximization of (present rational) prefer-

[81] This is the climax of Parfit's argument against the self-interest theory of rationality, *S*. Parfit 1984, 317–318.

[82] Parfit 1984, 312–313, 345–347.

[83] See page 77.

ences, and to do this we do not need a concept of a person providing historical continuity. From the point of view of Parfit's normative theory, there is no need for a wide, or "thick" concept of a person, but the kind defined by *Relation R* is sufficient, and the perspective in which we have to see morality, according to Parfit, is not the perspective of *S*, one's whole life, but that of each of one's present, co-existing preferences.

Our second question concerns the metaethical status of the *Relation R*. Parfit thinks that his reductionism is a true view of ourselves. Hence, it follows that we must abandon the self-interest theory of rationality, for it includes a mistaken view of our nature as persons. It is noteworthy that he does not abandon the aim, i.e., desire-satisfaction as the basis for intentional action, but only the perspective in which this aim must be seen. In doing this, Parfit seems to regard reductionism as a rational naturalistic foundation of the principle of universalizability in ethics: since reductionism is the true view of ourselves, our moral codes must require the same treatment for ourselves, for our future selves, and for other people, whether of not they are our contemporaries or future neighbours. Reductionism offers a formal, non-moral account of the good. We must, according to Parfit's utilitarian approach, concentrate on people's preferences and on their satisfaction. People care about each other's preferences and about their satisfaction because they are naturally benevolent. Benevolence is a natural sentiment, and as such no less rational than self-concern. Because we deal with ethics, we need criteria for cases in which people's preferences conflict. Personal identity is not what matters for what we are, and consequently, it cannot be given any weight in moral deliberation. Parfit's concept of a person matches with his utilitarian moral theory; features which are essential for determining the morally relevant and normative conclusions also play a central role in his concept of a person.

Whatever Parfit claims, his Reductionism does not provide us with a new basis for rationality or for ethics. Parfit's main argument for a new morality is that we are less closely connected with ourselves than we commonly think we are, and for this reason, it is not rational to care more for ourselves. But this not an argument in support of the rationality

of caring more for others. It is not even a motivating fact for a more unselfish way of living; for why should I care more for others just because I am told to care less for myself.[84]

Parfit't view becomes even more unintelligible if we start to think what it would actually require. If we take reductionism seriously how will it affect what we do, or should we say, what a person in general does? There are two possibilities. First, one will be concerned about how things go for a person, i.e., of the quality of sets of co-existing experiences, but one does not care "whose" sets of experiences these are. Is this intelligible? Could we speak of making plans and setting goals? Could we continue to speak of intentional actions, or would this convention of speech and thinking also have to be revised? It is difficult to answer these questions because the requirement of adopting no one's point of view to direct one's reasoning seems confused. The second possible way to interpret the consequences of Reductionism is to say that we no longer have a reason — in the old, non-reductionist sense — to care much for anyone's life, because there are actually no such thing as a person, or a person's life in the sense we always thought there was. This would, indeed, make a difference to how we live our lives, whatever that could mean, but this would be a change for the worse. To conclude, even if Parfit's reductionism were "the true view of ourselves" this would neither provide us with a better basis for ethics, nor help us to combine the goals of morality and rationality.

[84] Wolf (1986, 249) and Glover (1988, 105) make a similar point.

1.4. The utilitarian person

In the beginning of this chapter utilitarian theories were characterized as models whose dominant idea was that the aim of morality is to maximize the non-moral good of human happiness or welfare, defined as some sort of desire-satisfaction. The task in this chapter was to analyze the concept of a person in three utilitarian theories, and, if possible, to outline the utilitarian conception of a person. But is there such a conception? On the basis of the preceding analysis it seems justified to say that each of the three theories we have examined employ a conception which can be called the *utilitarian person*.

The analysis has shown that despite many differences, the three theories share some central features. All three theories see the heart of intentionality in the desiring agent of the practical syllogism. That something is desired means that people have interests and preferences. When these preferences concern other people's interests and preferences the situation becomes morally relevant. A natural agent becomes a moral subject when her intentional action affects other intentional agents' interests and preferences, or in other words, when her action places someone in the position of a moral object. The natural and the moral agent are, however, fundamentally similar. This fact reflects the nature of utilitarianism as a moral theory: the distinction between the moral and the non-moral is not sharp but the moral is the good or qualified natural.

The analysis also showed that the definition of the morally relevant in these theories squares with their definition of the moral person. Features which constitute a moral situation are the same as those that constitute the moral person. Interestingly, all the examined theories accept the definition of the morally relevant, and thus also the description of the moral person, as a self-evident given without problematizing the possible normative commitments it might involve. But we can say that the definition of the morally relevant has normative significance in the sense that one recognizes the morally relevant situation through this theoretical description and that this description is given for the purpose of directing

a person's action in any moral situation.

Although the utilitarian moral person is essentially defined through her desires, and through the way their satisfaction is understood the theories differ in their treatment of the desires in the moral realm. Different versions of utilitarianism represent diverging methods and grounds for prioritizing one desire over another, or for qualifying different desires. In Brandt's theory, the qualificatory task is given to the procedure of cognitive psychotherapy which both qualifies and justifies a morally relevant desire. Hare presents a twofold method for selecting among the desires. The intuitive level of moral reasoning disqualifies most of the desires conventional morality deems immoral. The more complicated situations which involve conflicting desires and irreconcilable moral rules are evaluated by means of critical thinking. Here the dispute between desires is settled on formal grounds. To achieve the outcome Hare claims to reach with his theory, it is necessary, however, to accept the presupposition that the desires of all human beings are more or less uniform. Parfit for his part does not clearly represent any distinct model of utilitarianism but maintains instead that there are several possible ways of treating the desires. It is clear, however, that in his theory, too, desires essentially constitute the person.

What is significant, and at its clearest the idea is present in Parfit, is that there is no need for a conception of a continuous person in these theories. The scope within which we see the person is temporally narrow; there is no need for a framework of a life, or of a developing moral character. The theories focus on people's desires among which the present ones are preferred. The crudest form of this understanding of a person is a version that reduces the concept of a person to a set of (mutually consistent) present desires. This formulation does not exclude the possibility of the desires being of "high" moral worth; the three theorists of this chapter all join the British tradition of moral philosophy which presupposes that people naturally have a tendency to be benevolent and to show regard for others.

In Brandt's theory the moral person is constituted upon the set of mutually consistent desires qualified by cognitive psychotherapy. The set

of one's present desires has a very central role in the theory. There is actually no need for a morally substantive concept of a person apart from, or over and above the person's present qualified desires, for only the "therapeutized" desires define the morally relevant.

The lack of any concept of a temporally continuous person is still more remarkable in Parfit's theory. It is one of Parfit's explicit aims to abandon, even to abolish, the substantive concept of a person. In his view, not only is there no need for such a conception in ethics, but even to cherish it is morally damaging. We can better understand Parfit's view against the central role he gives to desires in his theory. When we express the moral relevance of a situation we describe it by saying that it involves human agents who have differing and (possibly) conflicting desires of different strengths. Furthermore, the actual desires of the people involved in a certain situation provide the necessary information we need for our moral decision-making. When any morally relevant situation is conceived in these terms, a substantive concept of a person, involving a "further fact", becomes redundant; there is simply no use for such a notion. It is characteristic of Parfit's view that the *Relation R* which replaces the traditional concept of a person is applied not only as a theoretical device for an appropriate description of the morally relevant, but also as a corrective for practical moral reasoning. According to Parfit, we must namely abandon our traditional substantive concept of a person because this understanding distorts our moral thinking in a way that is, in addition to its mistaken factual basis, morally harmful. We have thus both factual and moral grounds for adopting the view of the *Relation R* as sufficient and necessary for the correct view of our lives and morality in general.

In Hare's theory, too, the correct description of the morally relevant features is of normative importance. If we understand the nature of moral statements and what they involve according to the theory, we can no longer support solutions that are morally blameworthy. Despite the central role given to the logical analysis of the moral language in Hare's theory, it does not have an effect on the way the morally relevant and the concept of a person are defined. As in the other two utilitarian theories

we have examined, the Harean moral person is defined through her desires, both as a moral subject and as a moral object.

The preceding study has shown that there is a tendency in the utilitarian theories for the moral point of view to take precedence. The concept of maximizing, which is central to these theories, is not a specifically moral notion. A rational intentional agent aims at maximizing the effects of her action, whatever she has set as her aim. Accordingly, the maximizing concept of morality does not impose any limits on moral activity but just as the needs of the utilitarian person — as a "consumer" of goods that satisfy her desires — are limitless, so are the demands that the utilitarian moral theory makes on her as a moral subject. Every intentional action is a possible candidate for moral action because the utilitarian good is a non-moral good and the distinction between the moral and the non-moral remains indefinite: everything that affects the interests of others is morally relevant. This has the consequence that the utilitarian moral person tends to replace the natural person of unqualified intentional action. I have called this feature of utilitarianism "moral imperialism".

The "imperialistic" nature of utilitarianism becomes understandable given that these theories combine morality with rationality. Brandt's explicit goal is to design a moral theory which reduces traditional moral questions to questions about rationality. For Parfit, a unified theory combining moral and rational reasons for action solves the difficulties he sees as plaguing modern moral philosophy. Hare, for his part, claims to have established moral thinking on a logically valid basis, which is not merely compatible with rational reasoning but also an integral part of it.

2. *Social contract and deontological theories*

The present chapter introduces three theories: John Rawls' famous contractarian theory on social justice; David Gauthier's revision of Rawls basic ideas in the form of "morals by agreement"; and finally Alan Gewirth's strongly rationalist moral theory, which attempts to find a rationally acceptable and logically necessary grounding for morality and to build a normative moral theory on that basis. In one particular Rawls' theory and Gewirth's model deviate strongly from each other. Rawls does not actually present an ordinary moral theory but a theory on social justice, which does not ground morality but presupposes it as an institution; it is a necessary condition for the kind of theory of justice Rawls wishes to develop. Gewirth for his part, adopts what could be called a traditional project in moral philosophy for he claims to have laid down a rational basis for the institution of morality itself. Where Rawls' interests focuses on a viable normative theory of justice, Gewirth starts off from a non-moral conceptual analysis which he designates to serve a justifiable normative theory. Gauthier's project is somewhere in the middle: he wishes to prove that morality can be predicated on rational grounds and, additionally, that a normative morality can be built on this rational basis.

What all these three theories have in common, though, and what makes it legitimate to present them together in the same chapter is the role these theories give to intentional action and its necessary constituents as a part of morality. They all share the view that intentional action becomes impossible without certain basic conditions. A rational agent wishes to secure herself at least the fulfilment of these necessary conditions, irrespective of what else she wishes to have. Morality is a special case of intentionality; and we enter the moral realm when our actions

affect other people's interests and when the actions of others have an effect on us. A rational agent, in considering the ends and means of her action, aims at securing those conditions of her action that are necessary for her whatever her ends are. A moral code arises when people recognize the mutual need of such prerequisites irrespective of their individual goals. This recognition gains in solidity through an analysis of the conditions of a rationally acceptable *contract*.

The theories of Gewirth, Gauthier and Rawls represent a view according to which morality is, above all, a *restriction* to what agents would otherwise do and not a positive means for human flourishing. This implies that morality has only a limited task: its task is primarily negative as it inhibits actions violating the conditions of rational agency which belong to all people. Morality is only secondarily exhortative in the sense that it would cause people to produce well-being and happiness. This shows that these theories accept as an implicit starting point a certain view of human beings. Human beings are taken as naturally selfish rather than benevolent. Rational agents have their own straightforward interest as their instant goal if the requirements of morality do not inhibit this.

2.1. John Rawls — personhood under given conditions

A Theory of Justice has made John Rawls one of the most prominent social theorists of the twentieth century.[1] Although Rawls' ideas do not so much form a moral theory as a concept of social justice they are worth examining even in the context of this study. In explicating the concept of a moral person Rawls employs in *A Theory of Justice*, I shall concentrate on two points. First, I examine Rawls' starting point and pay attention to certain features in his methodology. Second, I analyze the conception of the *original position* and related themes.

2.1.1. *JUSTICE AS FAIRNESS*

Rawls' purpose is to revive the social contract tradition and introduce it as an alternative to utilitarian and intuitionist moral theories which have dominated the discussion in moral philosophy.[2] Rawls is guided in his theory by the basic intuitive idea that justice is the most important fea-

---

[1] Rawls has published a renewed version of his theory: *Political Liberalism*. Columbia University Press. New York, 1993. This is a revised version of Rawls' central idea of justice as fairness. The main difference between the two versions of the theory lies in the concept of a well-ordered society. Unlike his *A Theory of Justice* the new book does not presuppose that the stability of a well-ordered society is based on shared moral beliefs but on its political conception of justice instead; see RAWLS 1993, xv–xvii. As my thesis is not an examination of Rawls' theory but a study of different concepts of a moral person *A Theory of Justice* offers a sufficent basis for the present work. Moreover, the development of Rawls' idea does not affect the concept of a person embedded in the theory: the basic concept remains unaltered. In the following presentation I will concentrate on analyzing *A Theory of Justice* and only occasionally refer to *Political Liberalism*. *A Theory of Justice* has been a source of inspiration for numerous writers who have both criticized and developed Rawls' ideas. In the present study I do not make use of this very extensive secondary literature unless it directly concerns the theme of this thesis.

ture of all social institutions; this is an intuition people universally share.[3] Accordingly, Rawls' theory justifies neither the institution of morality nor the basic principle of justice but starts off "from the middle"; it presupposes the concept of justice as a given prerequisite of social life.[4] Rawls' contractarian theory, which presents *justice as fairness,* derives from the deontological tradition.[5]

Despite his criticism of utilitarianism and intuitionism, Rawls' work shares some fundamental features with these approaches. Intuitionism is right in so far as it acknowledges the meaning of intuitions for moral theory, for there is nothing necessarily irrational in appealing to intuition in

---

[2] This was the situation when Rawls published his book, but the theoretical variety of moral philosophy has become considerably wider since the beginning of the 1970s. Rawls succeeded in his attempt to revive the social contract tradition, and it is now a major trend in ethical theory. "In presenting justice as fairness I shall contrast it with utilitarianism. I do this for various reasons, partly as an expository device, partly because the several variants of the utilitarian view have long dominated our philosophical tradition and continue to do so. And this dominance has been maintained despite the persistent misgivings that utilitarianism so easily arouses. The explanation for this peculiar state of affairs lies, I believe, in the fact that no constructive alternative theory has been advanced which has the comparable virtue of clarity and system and which at the same time allays these doubts. Intuitionism is not constructive, perfectionism is unacceptable. My conjecture is that the contract doctrine properly worked out can fill this gap. I think justice as fairness an endeavor in this direction." RAWLS 1972, 52.

[3] "It has seemed to many philosophers, and it appears to be supported by the convictions of common sense, that we distinguish as a matter of principle between the claims of liberty and right on the one hand and the desirability of increasing aggregate social welfare on the other; and that we give a certain priority, if not absolute weight, to the former. Each member of society is thought to have an inviolability founded on justice, or, as some say, on natural right, which even the welfare of every one else cannot override." RAWLS 1972, 27–28. See also RAWLS 1972, 3–4.

[4] "Let us assume, to fix ideas, that a society is a more or less self-sufficient association of persons who in their relations to one another recognize certain rules of conduct as binding and who for the most part act in accordance with them. Suppose further that these rules specify a system of co-operation designed to advance the good of those taking part in it." RAWLS 1972, 4; and: "For given the circumstances of the original position, the symmetry of everyone's relation to each other, this initial situation is fair between individuals as moral persons, that is, as rational beings with their own ends and capable, I shall assume, of a sense of justice." RAWLS 1972, 12. There is, nevertheless, also a justificatory aspect in Rawls' theory: starting off from an intuitive concept of justice he aims at developing and refining it as well as at giving it a grounding; see RAWLS 1972, 4.

[5] RAWLS 1972, vii–viii.

moral philosophy.[6] The mistake of intuitionism is rather that it represents the thought that there is an irreducible collection of first moral principles but no set of higher-order constructive criteria for justifying or ranking them. This line of thought easily leads to ethical pluralism, however.[7] Justice as fairness makes use of ethical intuitions, too, but contrary to intuitionism it includes a procedure for arranging them.[8]

Utilitarianism is intuitively appealing while it defines the good as maximization of satisfaction. The concept of maximization helps to connect morality with rationality, because rationality embodies the idea of maximization, and it is convenient to think that in the moral realm one maximizes the good. Rawls adopts the version of this notion, which defines the good as satisfaction of rational desire.[9] Utilitarianism as a whole is, however, not acceptable as a moral theory, because it prioritizes the good over the right, a position which violates our fundamental moral intuitions.[10] In contrast to the utilitarian approach, the good must always

---

[6] "We must recognize the possibility that there is no way to get beyond a plurality of principles. No doubt any conception of justice will have to rely on intuition to some degree. Nevertheless, we should do what we can to reduce the direct appeal to our considered judgments. For if men balance final principles differently, as presumably they often do, then their conceptions of justice are different." Rawls 1972, 42. "[...] there is nothing necessarily irrational in the appeal to intuition to settle questions of priority [between different moral principles]." Rawls 1972, 41.

[7] "I shall think of intuitionism in a more general way than is customary: namely, as the doctrine that there is an irreducible family of first principles which have to be weighed against one another by asking ourselves which balance, in our considered judgment, is the most just. Once we reach a certain level of generality, the intuitionist maintains that there exist no higher-order constructive criteria for determining the proper emphasis for the competing principles of justice." Rawls 1972, 34.

[8] "The only way therefore to dispute intuitionism is to set forth the recognizably ethical criteria that account for the weights which, in our considered judgments, we think appropriate to give to the plurality of principles. A refutation of intuitionism consists in presenting the sort of constructive criteria that are said not to exist. To be sure, the notion of a recognizably ethical principle is vague, although it is easy to give many examples drawn from tradition and common sense. But it is pointless to discuss this matter in the abstract. The intuitionist and his critic will have to settle this question once the latter has put forward his more systematic account." Rawls 1972, 39.

[9] Rawls 1972, 30.

[10] For a more detailed exposition of Rawls' criticism against utilitarianism, see Rawls 1972, 25–31.

have a subordinate role in relation to the deontologically defined right.[11] In the fully developed theory of *justice as fairness*, the two must, nevertheless, be fully compatible with each other so that realizing justice as fairness accords with goodness as rationality.

2.1.2. THE *REFLECTIVE EQUILIBRIUM*

Rawls develops his theory through a deductive procedure called *reflective equilibrium*. The position he gives to intuitions in the theory can be explicated in the following way. First, loosely defined moral intuitions are taken to form the initial *a priori* first principles for a tentative conception of justice. These intuitions state that the necessary condition for calling any societal organization or procedure just is that it secures the basic equality of human beings against the effect of contingent differences. Thus, any conception of justice must be refuted if it does not satisfy this intuitively necessary condition of equality.[12] We can describe this use of intuitions as *substantive*, which means that intuitions are used for determining the basic normative principle of justice.

Besides this use, Rawls employs intuitions in another, *methodological* fashion. Rawls presupposes that there are, in addition to substantive

[11] "[...] utilitarianism is a teleological theory whereas justice as fairness is not. By definition then, the latter is a deontological theory, one that either does not specify the good independently from the right, or does not interpret the right as maximizing the good. [...] Justice as fairness is a deontological theory in the second way. For if it is assumed that the persons in the original position would choose a principle of equal liberty and restrict economic and social inequalities to those in everyone's interest, there is no reason to think that just institutions will maximize the good. [...] The question of attaining the greatest net balance of satisfaction never arises in justice as fairness; this maximum principle is not used at all." RAWLS 1972, 30. See also RAWLS 1972, 31–32.

[12] RAWLS 1972, 20, 22, 49. Interestingly, Rawls does not purport to prove that his claim regarding the universality of the intuition concerning justice is correct. He does not even offer any evidence for his claim. The starting point of moral thinking simply is the priority of justice over other moral values, and the core of justice is a guarantee against contingent differences.

moral intuitions, intuitions which define the formal features of normative moral principles. This means that there are certain formal criteria which principles of justice have to fulfil so as to conform with our intuitions. These conditions can be derived from the role that principles of justice have in ordering human communities.[13] We can also say that these intuitions define the conditions of human life which we intuitively think should be regulated by the institution of morality.[14] Let us consider some of these methodological intuitions.

Rawls maintains, without giving grounds for his opinion, that morality essentially sets negative *restrictions* on people's behaviour rather than positive principles for attaining some desired state of affairs. Furthermore, he requires that the principles of justice must conform with such formal conditions as *generality*, *universality*, *publicity* and *finality*. In addition, they must establish an ordering by which conflicting claims can be given a uniform *order of priority*.[15] In Rawls' view, these formal criteria do not rule out any of the traditional conceptions of justice, thus preserving these variants as possible candidates to be considered in more detail. In contrast to these, the crudest forms of egoism are disqualified: they do not satisfy the formal criteria which our intuitions establish for moral principles.[16] This view implies that we do not have to oppose these forms of egoism within ethics, for their "defeat" is secured even before

[13] "The propriety of these formal conditions is derived from the task of principles of right in adjusting the claims that persons make on their institutions and one another. If the principles of justice are to play their role, that of assigning basic rights and duties and determining the division of advantages, these requirements are natural enough." Rawls 1972. 131.

[14] Rawls maintains that we all intuitively think that principles of justice must hinder anyone from taking advantage of her natural fortune or social circumstances, as well as making it impossible for everybody to tailor principles governing justice to match her own situation, inclinations, aspirations and the like. Consequently, it is a necessary condition for any set of principles ruling justice that it rules out inequality arising from these sources. This means that the conditions which regulate the formulation of the basic principles must be designed so as to produce such principles which fulfil these necessary criteria of justice. Rawls 1972, 18–19.

[15] Rawls 1972, 131–135. "Taken together, then, these conditions on conceptions of right come to this: a conception of right is a set of principles, general in form and universal in application, that is to be publicly recognized as a final court of appeal for ordering the conflicting claims of moral persons." Rawls 1972, 135.

we enter the moral realm, on purely formal grounds.

Nussbaum criticizes Rawls for adopting these formal criteria. In her view, such criteria necessarily rule out various aspects of human experience, and thus, ignore them as ethically irrelevant; most notably the particular and contingent features in human life are unduly disregarded. Consequently, the reflective equilibrium cannot be complete as an ethical method, for it excludes, against Rawls' explicit aim, morally significant approaches from appropriate concern, and leaves large parts of human life without attention.[17] Rawls could answer this criticism by maintaining that the restrictiveness of these formal criteria, which Nussbaum regards as a weakness, is actually the strength of his theory. Justice as fairness joins the deontological tradition that determines the criteria of the morally relevant from, so to say, outside, irrespective of what people in a particular situation prefer or want. Consequently, it becomes unambiguous to say what is morally relevant and what is not. Or, Rawls could defend his approach by maintaining that his theory is not primarily a moral theory but a theory of social justice which presupposes the existence of morality and the importance of a rich variety of human values, one of the most central of which is self-respect.[18]

These two kinds of intuitions, those regarding normative moral principles and those concerning their formal criteria, form the two parts between which the dialectical process of Rawls' method, the reflective equilibrium, takes place. I adopt the following interpretation of this pro-

---

[16] "Now by themselves the five conditions exclude none of the traditional conceptions of justice. It should be noted, however, that they do rule out [...] variants of egoism. The generality condition eliminates both first-person dictatorship and the free-rider forms, since in each case a proper name, or pronoun, or a rigged definite description is needed, either to single out the dictator or to characterize the free-rider. Generality does not, however, exclude general egoism, for each person is allowed to do whatever, in his judgment, is most likely to further his own aims. The principle here can clearly be expressed in a perfectly general way. It is the ordering condition which renders general egoism inadmissible, for if everyone is authorized to advance his aims as he pleases, or if everyone ought to advance his own interests, competing claims are not ranked at all and the outcome is determined by force and cunning." RAWLS 1972, 135–136.

[17] See footnote 28 on page 198.

[18] See, RAWLS 1972, 433–434.

cedure: on the one side, there are the intuitive *a priori* restrictions, or what we "naturally" believe justice should prevent happening, on the other side, there are tentative criteria for deducing the normative principles of justice. The outcome of the deliberative exchange in the reflective equilibrium is twofold. First, the intuitive first principles of justice develop into a reflected and justifiable normative theory concerning social justice. Second, the intuitions, which express the formal criteria that acceptable moral principles must fulfil, turn into a fully defined concept of the morally relevant from the perspective of social justice. Using the terminology of Rawls' theory, these intuitions will form the notion of an *original position*. In the final stage of the theory these two components are fully compatible with each other: the normative theory corresponds with the formal criteria defined by the concept of an original position, and the original position contains the premises for the deduction of the normative principles. All in all, reflective equilibrium is an ongoing exchange between these two levels of intuitions which constantly mould and correct each other. This procedure is applied in both directions as long as an equilibrium between the definition of the original position and the principles of justice has been achieved. When this stage has been reached the theory of justice as fairness is fully developed.[19]

One of the advantages of the procedure is, according to Rawls, that it takes people's moral intuitions into account both in the sense that intuitions form the initial starting point of the theory, and in the sense that the final outcome of the procedure is intuitively acceptable. A further

[19] "In searching for the most favored description of this situation we work from both ends. We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield a significant set of principles. If not, we look for further premises equally reasonable. But if so, and these principles match our considered convictions of justice, then so far well and good. But presumably there will be discrepancies. In this case we have a choice. We can either modify the account of the initial situation or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth sometimes altering the conditions of the contractual circumstances, at others with drawing our judgments and conforming them to principle. I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium." Rawls 1972, 20.

advantage of reflective equilibrium is that it makes the theory of justice rationally justifiable. The reflective nature of the procedure requires that every suggestion is subject to constraints of rationality, given the intuitively acceptable initial assumptions. The method finally produces a moral theory which is an intuitively plausible and a rationally justifiable *contract* between rational agents, a contract which every rational agent can accept as her own.[20]

As was noticed earlier,[21] Rawls' theory lacks an explicit metaethical justification of morality. The exposition of reflective equilibrium can, however, be interpreted as such. Namely, Rawls presents reflective equilibrium as an ethical method, but readily admits that he does not apply it for developing his own theory.[22] Why then develop and present it at all? What can the purpose of the method be? In short, its object is to justify the validity of the principles of justice by explicating and developing the intuitive sense of justice, and by providing them with a considered grounding.[23]

---

[20] "We shall want to say that certain principles of justice are justified because they would be agreed to in an initial situation of equality. I have emphasized that this original position is purely hypothetical. It is natural to ask why, if this agreement is never actually entered into, we should take any interest in these principles, moral or otherwise. The answer is that the conditions embodied in the description of the original position are ones that we do in fact accept. Or if we do not, then perhaps we can be persuaded to do so by philosophical reflection. Each aspect of the contractual situation can be given supporting grounds. Thus what we shall do is to collect together into one conception a number of conditions on principles that we are ready upon due consideration to recognize as reasonable. These constraints express what we are prepared to regard as limits on fair terms of social cooperation." RAWLS 1972, 21.

[21] See page 106.

[22] RAWLS 1972, 21.

[23] "These constraints [set by the original position] express what we are prepared to regard as limits on fair terms of social cooperation. One way to look at the idea of the original position, therefore, is to see it as an expository device which sums up the meaning of these conditions and helps us to extract their consequences. On the other hand, this conception is also an intuitive notion that suggests its own elaboration, so that led on by it we are drawn to define more clearly the standpoint from which we can best interpret moral relationships. We need a conception that enables us to envision our objective from afar: the intuitive notion of the original position is to do this for us." RAWLS 1972, 21–22.

2.1.3. The original position as an explication of the moral person

According to Rawls' definition, the original position determines the conditions under which the agreements reached are fair.[24] As outlined in the theory this position is completely hypothetical and conceptual; there has never been and there will never be a "conference" described by the original situation, instead it can represent any moment of time.[25] How should the original position, then, be deconstructed? According to my interpretation of the original position, it is an explication of what we intuitively regard as morally relevant. To understand what further significance this notion has, let us examine it closer.

Rawls explicates what he means by the original position by defining the characteristics and qualities of its *participants*. The participants of the original position are rational agents who act to pursue their own ends. The nature of their rationality is instrumental: they will choose the most effective means to attain their individually defined ends. In a world of scarce rather than abundant resources instrumental rationality makes them seek co-operation with others, even if this involves confining their own liberty; in other words, they acknowledge the conditions of justice. The ends of the participants, or their *good*, are determined by their informed desire. Furthermore, the participants in the original position are moral: they have a sense of justice and they keep their promises and

[24] "Just as each person must decide by rational reflection what constitutes his good, that is, the system of ends which it is rational for him to pursue, so a group of persons must decide once and for all what is to count among them as just and unjust. The choice which rational men would make in this hypothetical situation of equal liberty, [...], determines the principles of justice." Rawls 1972 11–12. "The original position is, one might say, the appropriate initial status quo, and thus the fundamental agreements reached in it are fair. This explains the propriety of the name "justice as fairness": it conveys the idea that the principles of justice are agreed to in an initial situation that is fair." Rawls 1972, 12.

[25] "In justice as fairness the original position of equality corresponds to the state of nature in the traditional theory of the social contract. This original position is not, of course, thought of as an actual historical state of affairs, much less as a primitive condition of culture. It is understood as a purely hypothetical situation characterized so as to lead to a certain conception of justice." Rawls 1972, 12.

comply with the agreements they have reached with others. Each participant is a solitary individual who pursues her own particular goals but who acknowledges the restrictions of morality in this pursuit. The participants neither cherish interest in other persons' good, nor suffer from envy. They do, nevertheless, know that they are not alone in the original position, and that the other occupants of the initial state are similar to themselves: rational disinterested agents who, adopting the moral constraint, strive for their particular goals.[26]

What status does this concept of a participant of the original position have in Rawls' theory? The original position defines what is morally relevant, and consequently, the participant of the original position gives the characteristics of a moral person in the theory.[27] We notice here that the concept of a person is defined primarily as an agent, endeavouring to actualize her own goals, and only secondarily as a moral agent. She is moral in a negative, but not in a positive sense: the agent refrains from inhibiting other agents as they pursue their own good, when each one is secured a similar right to pursue their good by a mutually reached agreement, but she does not seek to further other persons' good. Although Rawls presupposes that the person acknowledges morality and acts in a moral realm, her compliance to morality seems, in the last instance, to be justified by her instrumental rationality. She accepts moral constraints because she knows that, other agents having a similar concept of rationality, these will guarantee her the conditions which are necessary for the pursuit of her own, individual goals, because the same constraints will

[26] RAWLS 1972, 4, 12, 94–95, 144–145. RAWLS (1993, 29–35) presents a political concept of a person which accords with the interpretation I have given here. According to Rawls, people have a public and a private identity. The private identity is based on their personal conception of the good. It and thus people's conception of themselves can change, even radically, but this does not affect their public identity which determines their co-operation with others in the political sphere. The theory does not determine the conception of the good underlying one's private identity as long as it is compatible with the public notion of justice.

[27] A standard, more pragmatic interpretation of the original position is that it represents a testcase for solving moral dilemmas. When facing a moral problem, one should imagine oneself as a participant of the original position trying to solve this dilemma. The conclusion to which one then comes is what one, in one's real life, should do. See KUKATHAS & PETTIT, 1990, 18–26.

also guarantee similar conditions for other persons. The person is, thus, an ethical "minimalist", not a supererogatory or even a benevolent agent.

One can object to this interpretation by maintaining that as Rawls' theory presupposes morality and the parties' compliance with it, it is not correct to claim that morality is only given an instrumental status. The Rawlsian moral agent accepts the moral constraint because self-respect is one of her basic goods and because compliance with morality forms a central part of this self-respect.[28] It is, however, worth noticing that the normative assumptions Rawls constructs as presuppositions of his theory must conform with the formal requirements of justice as fairness. Consequently, self-respect and moral compliance, although presupposed by the theory, must later be formulated so as to square with justice as fairness and the formal conditions which determine it.[29] Furthermore, Rawls defines the qualities of self-respect and moral goodness as properties which it is rational to want in persons, whether in oneself or in others, irrespective of the person's particular characteristics.[30] Against this background, we can say that morality in justice as fairness is not instrumental, but that the motives the participants of the original position have for adopting this morality arise from instrumental rationality. The original position produces an altruistic morality out of the parties' egoistic interests. This does not mean, however, that the parties will necessarily have to be egoistic, they can turn out to be as benevolent and altruistic as possible. What the procedure does require is that the concept of a *moral*

[28] Rawls regards self-respect as one of the basic goods because: " it includes a person's sense of his own value, his secure conviction that his conception of his good, his plan of life, is worth carrying out. And second, self-respect implies a confidence in one's ability, so far as it is within one's power, to fulfill one's intentions. [...] It is clear then why self-respect is a primary good. Without it nothing may seem worth doing, or if some things have value for us, we lack the will to strive for them." Rawls 1972, 440. Self-esteem makes rational agents protect the self-respect of others, for they understand that mutual respect is the basis of their own self-respect as social beings living in a society. This social grounding of self-respect also contributes to the rational agents' compliance with morality: a person can maintain her own self-respect only if the image others have of her and she has of herself is not impaired by moral shame; see, Rawls 1972, 440, 442, 445.

[29] See, Rawls 1972, 433–435.

[30] See, Rawls 1972, 433, 435

person, on the basis of which the contract is made, is a narrow one.[31]

Although the agent in the original position is already moral, there are two features which counteract morality: her disinterestedness in other people's good, and her instrumental conception of rationality. Consequently, she will attempt to get as much of the co-operational goods as possible with as few costs to herself as possible. This makes the agent liable of taking advantage of her specific, contingent features. Such partiality, or bias towards oneself, conflicts, nevertheless, with the intuitive conception of justice underlying Rawls' theory. According to this intuition, human persons are fundamentally equal with each other irrespective of their contingent differences.[32] For this reason, there has to be something in the theory that inhibits people's selfish bias and secures a just outcome. Rawls designs the concept of the *veil of ignorance* to serve this purpose in his theory. But before examining the veil of ignorance in more detail, it is useful to discuss the necessity of such a constricting concept.

Why does Rawls, in the first place, define the person in the original position in a way that it contains two mutually incompatible features, namely, that the person is a *moral* agent, and that she is *biased* in her own favour? First, it is not Rawls' aim to justify morality, and for this reason the institution of morality can be taken as a given in the theory. This initial conception is, nevertheless, not a substantive notion of morality, i.e.,

---

[31] See POGGE 1989, 94–95 for a similar interpretation. POGGE (1989, 86–94) criticizes SANDEL (1982) for falsely interpreting Rawls' concept of a person as a "deontological self". The deontological self is barren of all characteristics which make choice possible and it does not allow for a mutually shared self-understanding constitutive of the institutions of the community, something Sandel's communitarian theory would require. Sandel's mistake is that he does not recognize that the Rawlsian person is designated for social institutions based on a rational contract, and that this concept does not determine the characteristics of actual individuals even if they comply with the conditions of the contract.

[32] "Thus it seems reasonable and generally acceptable that no one should be advantaged or disadvantaged by natural fortune or social circumstances in the choice of principles. It also seems widely agreed that it should be impossible to tailor principles to the circumstances of one's own case. We should insure further that particular inclinations and aspirations, and persons' conceptions of their good do not affect the principles adopted." RAWLS 1972, 18.

one including specific normative rules, but only a formal concept, acknowledging the binding nature of moral rules when they have been established by some acceptable procedure. Thus, the agent is a moral person only in a formal sense: she recognizes the institution of morality, but does not thereby commit herself to any particular set of moral rules. In view of this, it is feasible to say that disinterestedness in other people's good and an instrumental concept of rationality have an impact parallel to the effect which derives from the formal nature of morality: they guarantee that the social agreement which the parties make is not grounded upon a substantive moral concept, but on formal criteria. The role of these formal criteria is to secure the universalizability of the contractual outcome: a concept of justice in the original position does not depend on an agreement concerning certain normative values, but on certain universally acceptable formal principles. This means, however, that the procedure does not necessarily guarantee impartiality, and for this reason the effects of self-bias must be nullified by the veil of ignorance.

The most important task of the veil of ignorance is to annul the effect of contingencies in the negotiation which establishes justice in order to prevent biased outcomes.[33] In the original position, when nothing has yet been agreed, the veil of ignorance covers all specific facts about both the individuals and the society for which the contract must be designed. The only information the parties are allowed to use in the

[33] "The aim is to rule out those principles that it would be rational to propose for acceptance, however little the chance of success, only if one knew certain things that are irrelevant from the standpoint of justice. [...] To represent the desired restrictions one imagines a situation in which everyone is deprived of this sort of information. One excludes the knowledge of those contingencies which sets men at odds and allows them to be guided by their prejudices. In this manner the veil of ignorance is arrived at in a natural way. This concept should cause no difficulty if we keep in mind the constraints on arguments that it is meant to express. At any time we can enter the original position, so to speak, simply by following a certain procedure, namely, by arguing for principles of justice in accordance with these restrictions." Rawls 1972, 18–19. "Somehow we must nullify the effects of specific contingencies which put men at odds and tempt them to exploit social and natural circumstances to their own advantage. Now in order to do this I assume that the parties are situated behind a veil of ignorance. They do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations." Rawls 1972, 136–137.

process covers the basic knowledge of human society and its functioning, social organizations, and both social and individual psychology. The veil of ignorance conceals the parties' position in society: their social class and status. They have no information about their natural abilities, talents, character, intelligence, physical strength, their lifeplans, or about their conception of the good. The parties are likewise ignorant of the generation they belong to. The contingent features of the particular society lie behind the veil of ignorance: the economical system, the political conditions, as well as the stage of civilisation and culture are unknown in the initial stage of the negotiation.[34]

What significance does it have for the theory that the veil of ignorance covers all this information? The role of the veil of ignorance is to curb the moral agent's bias in her own favour, and hence inhibit the parties from taking advantage of their distinct features in the original position. As we have noticed, the original position defines what is morally relevant; against this background, the facts which the veil of ignorance hide represent the morally irrelevant. Such information must, for that reason, be excluded from moral reasoning or ethical consideration. The facts, again, which the veil does not cover are properties having moral relevance. The information available to the agents at the initial stage is, then, either humanly universal or intuitively self-evident.

The contingent facts Rawls allows the parties to know as they enter the original position concern first, the conditions of justice, and second, certain psychological, sociological and economic facts of human beings and societies.[35] Rawls seems to presuppose that our psychological, sociological and economic knowledge forms a factual basis applicable to all

---

[34] "It is assumed then, that the parties do not know certain kinds of particular facts. First of all, no one knows his place in society, his class position of social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like. Nor, again, does anyone know his conception of the good, the particulars of his rational plan of life, or even the special features of his psychology such as his aversion to risk or liability to optimism or pessimism. More than this, I assume that the parties do not know the particular circumstances of their own society. That is, they do not know its economic or political situation, or the level of civilization and culture it has been able to achieve. The persons in the original position have no information as to which generation they belong." RAWLS 1972, 137.

human communities. He does not, however, explicate what he means by this knowledge. Similarly, he does not consider the possibility that the information could change or develop; what we now regard as a true factual description of the reality can later be refuted and replaced by something else.

The role given to the veil of ignorance displays one of Rawls' principal intuitionist presuppositions. For Rawls, one of the central tasks of justice is to nullify the effects of contingent differences, both innate and social. In his view, nobody has deserved their natural endowments. When a conception of justice is built on this presupposition it follows that these personal properties must not be treated as a person's private good but as a common resource. The original position must, thus, include a mechanism, i.e., veil of ignorance, for ruling out such variations.[36]

The basic idea in the original position is that the parties must reach an agreement under the conditions set by the veil of ignorance. Only moves which are rational in the light of these initial restrictions are accepted as steps towards the final agreement.[37] What would an agent under the conditions of the original position choose? According to Rawls, her choice would be directed by considerations common to all agents, and hence her decisions would represent the most central features of what is morally relevant. In the original position the agent would only choose what she chose regardless of whatever else she were to choose; in other words, she would select the necessary conditions for her being the kind of an

---

[35] "As far as possible, then, the only particular facts which the parties know is that their society is subject to the circumstances of justice and whatever this implies. It is taken for granted, however, that they know the general facts about human society. They understand political affairs and the principles of economic theory; they know the basis of social organization and the laws of human psychology. Indeed, the parties are presumed to know whatever general facts affect the choice of the principles of justice. There are no limitations on general information, that is, on general laws and theories, since conceptions of justice must be adjusted to the characteristics of the systems of social cooperation which they are to regulate, and there is no reason to rule out these facts." Rawls 1972, 137–138. To mention the conditions of justice in this connection is actually redundant: that they prevail is already included in Rawls' conception of the moral agent acting as a party in the original position. See page 114.

agent Rawls has defined an agent to be. She would decide on a set of principles which will guarantee all members of a society the largest possible amount of basic social goods, that is, *liberties and rights*.

Although no one in the original position has knowledge regarding her conception of the good, and the like, it is rational for everyone to be in possession of as many social goods as possible, in contrast to fewer ones. A rational person wants to enlarge her freedoms, prospects and rights as much as possible, and to guarantee for herself the opportunity to take advantage of different means for achieving her particular ends, whatever they turn out to be, when the veil of ignorance is lifted.[38] This, according to Rawls, means that the two principles which the parties in the original position would choose read as follows: "Each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others", and "Social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all".[39] Of these two the first is primary to the second one, liberty is a fundamental basic good,

---

[36] "Those who have been favored by nature, whoever they are, may gain from their good fortune only on terms that improve the situation of those who have lost out. The naturally advantaged are not to gain merely because they are more gifted, but only to cover the costs of training and education and for using their endowments in ways that help the less fortunate as well. No one deserves his greater natural capacity nor merits a more favorable starting place in society. But it does not follow that one should eliminate these distinctions. There is another way to deal with them. The basic structure can be arranged so that these contingencies work for the good of the least fortunate." RAWLS 1972, 101–102. See also RAWLS 1972, 104. In criticizing the liberal conception of justice Rawls maintains: "While the liberal conception seems clearly preferable to the system of natural liberty, intuitively it still appears defective. For one thing, even if it works to perfection in eliminating the influence of social contingencies, it still permits the distribution of wealth and income to be determined by the natural distribution of abilities and talents. Within the limits allowed by the background arrangements, distributive shares are decided by the outcome of the natural lottery; and this outcome is arbitrary from a moral perspective. There is not more reason to permit the distribution of income and wealth to be settled by the distribution of natural assets than by historical and social fortune." RAWLS 1972, 73–74.

[37] "To say that a certain conception of justice would be chosen in the original position is equivalent to saying that rational deliberation satisfying certain conditions and restrictions would reach a certain conclusion." RAWLS 1972, 138. See also RAWLS 1972, 12, 19.

because it is a prerequisite for all other human goods, and it may not be overridden by any other social good, nor be exchanged for anything else.[40]

What does the choice of the basic principles of justice involve? The conditions which direct this choice represent the morally relevant features of any situation. For what kind of a person do the parties choose the principles? How can we characterize the concept of a person explicable in the normative principles of Rawls' theory? Here *liberty* constitutes the moral person. Liberty, or freedom is a necessary condition for realizing any plan of life, and particular individual goods (defined as maximizing rational desire), whatever they happen to be. The task of the principles of justice is to regulate the conditions that direct each person when she aims at her own good: their task is to guarantee everyone as much as possible of the basic social goods every rational agent would want herself, no matter what else she would prefer, as prerequisites for any intentional human action. Although liberty receives such a central

---

[38] "As a first step, suppose that the basic structure of society distributes certain primary goods, that is, things that every rational man is presumed to want. These goods normally have a use whatever a person's rational plan of life. For simplicity, assume that the chief primary goods at the disposition of society are rights and liberties, powers and opportunities, income and wealth. These are the social primary goods." RAWLS 1972, 62. "Now the assumption is that though men's rational plans do have different final ends, they nevertheless all require for their execution certain primary goods, natural and social. Plans differ since individual abilities, circumstances, and wants differ; rational plans are adjusted to these contingencies. But whatever one's system of ends, primary goods are necessary means. Greater intelligence, wealth and opportunity, for example, allow a person to achieve ends he could not rationally contemplate otherwise. The expectations of representative men are, then, to be defined by the index of primary social goods available to them. While the persons in the original position do not know their conception of the good, they do know, I assume, that they prefer more rather than less primary goods. And this information is sufficient for them to know how to advance their interests in the initial situation." RAWLS 1972, 93.

[39] RAWLS 1972, 60.

[40] "These principles are to be arranged in a serial order with the first principle prior to the second. This ordering means that a departure from the institutions of equal liberty required by the first principle cannot be justified by, or compensated for, by greater social and economic advantages. The distribution of wealth and income, and the hierarchies of authority, must be consistent with both the liberties of equal citizenship and equality of opportunity." RAWLS 1972, 61.

position as a constituent of the Rawlsian moral person it is not just any kind of liberty. The second principle which regulates the distribution of inequalities namely modifies it to into a tool for the common good of the contractarian community.

To summarize the discussion, the description of the original position defines the morally relevant through the conception of a participant. What is significant in these participants is that the morally relevant is defined as a conglomerate of individuals pursuing their chosen goods through the most effective means with regard to their ends. Despite acknowledging the binding nature of moral rules, these agents accept the agreement above all, because doing this accords with their instrumental rationality. They know that other agents reason as they do, so making an agreement with them becomes a rational means for securing the necessary conditions of action for pursuing whatever one wants to pursue. The veil of ignorance then signifies those features one is not allowed to presume when one adopts the moral point of view. Making decisions as an occupant of the original position behind the veil of ignorance is tantamount to abandoning every consideration which has any personal meaning. Rawls' theory does not deny the existence and the importance of such personal aspirations, it just regards them as morally irrelevant.

2.2. David Gauthier — morals by agreement

David Gauthier's contractual theory is an attempt to justify morality, not to explain it as human institution. He joins the "Kantian" tradition which regards morality as a constraint for action in contrast to the "Aristotelian" which treats it as a model for good life. He aims at showing, starting off from non-moral premises, that it is rational to be moral. As such Gauthier's theory is part of a larger theory of (practical) rationality.[1] Gauthier develops his idea as follows: first, he presents a theory of action based on a maximizing concept of rationality. In this connection he uses a game-theoretical model for showing that it is rational to join co-operative ventures under certain conditions. The second part of the theory introduces a strategy model for rational co-operation. The strategy comprises of constrained maximization in which moral rules have the role of a constraint. Gauthier completes his theory with a discussion on the nature of morality as an end in itself in contrast to a mere means for other goals.

For the purpose of the present task, it is important to notice that each stage of the theory involves a concept of a person. The first part, dealing with the characteristics of rationality, presents a picture of a *natural man*, a concept used to explain the basic features of human intentionality. The second part introduces an *economic man*, a concept which explains the rationality model of constrained maximization. The third level finally brings the *liberal individual* on stage presenting Gauthier's view of what authentic moral personhood involves.[2] I will present the three levels of

[1] GAUTHIER 1988, 2–3. Gauthier joins the modern contractual tradition initiated by Rawls, but he claims to exhaust the possibilities of the contractarian model better than Rawls; his aim is both to give a rational justification to morality and to construct a normtive theory of justice. GAUTHIER 1988, 5.

[2] Gauthier explicitly uses the three different concepts of a person in developing his thoughts although he does not connect them as straightforwardly with the three levels of his theory as I have done.

Gauthier's theory by way of background to his concept of a moral person.

2.2.1. THE PORTRAIT OF A *NATURAL MAN*

There are two standard cases for examining human action: first, situations in which the agent acts solitarily and in which the outcome depends only on her decisions and actions, and second, cases where other agents, their respective decisions and actions have an impact on the situation. Gauthier first drafts a model for evaluating an agent's choice in a solitary situation, i.e., a *parametric* choice. He then continues by examining on what conditions and with what modifications the model of solitary action can be expanded to cover the *strategic* choices that are needed in interactive situations.[3]

In explaining human action Gauthier stresses the role of the agent's interest: *individual preference* plays a central part in action. Human action springs from the attempt to realize one's goals, and people's goals are things which they, in oneway or another, prefer.[4] This means that a rational agent attempts to maximize the fulfilment of her preferences; practical rationality is the *maximization of utility*. Here Gauthier joins the Humean tradition in which desire and volition dictate the objects of

---

[3] "Our present focus is on *parametric* choice, in which the actor takes his behaviour to be the sole variable in a fixed environment. In parametric choice the actor regards himself as the sole centre of action. Interaction involves *strategic* choice, in which the actor takes his behaviour to be but one variable among others, so that his choice must be responsive to his expectations of others' choices, while their choices are similarly responsive to their expectations." GAUTHIER 1988, 21.

[4] "Let us suppose it agreed that there is a connection between reason and interest — or advantage, benefit, preference, satisfaction, or individual utility, since the differences among these, important in other contexts, do not affect the present discussion. Let it further be agreed that in so far as the interests of others are not affected, a person acts rationally if and only if she seeks her greatest interest or benefit." GAUTHIER 1988, 6–7. See also GAUTHIER 1988, 22.

one's choice, whereas the role of reason is just to choose the best means for attaining the desired end. This has an effect on how morality is understood: the content of the appetites does not belong to the moral realm since reason is inappropriate for criticizing them. Appetites simply come into existence according their particular laws.[5]

In introducing his theory, Gauthier explicitly avoids giving any substantial definitions to the concepts he uses for explaining human action. He defines utility simply as a formal measure for preference.[6] The same applies to the conception of a preference: the theory is confined to giving certain *formal criteria* for outlining the rationality of preferences. These criteria are designed to determine the degree of consistency in an agent's preferences and the relative probability attached to different preferential outcomes.[7] By the same token, the concept of practical reason is strictly instrumental, t its aim being "the maximal fulfilment of our present considered preferences, where consideration extends to all future effects in so far as we may now foresee them."[8]

These considerations determine the rationality of intentional action when the agent acts alone and when her goal-achievement is independ-

---

[5] Gauthier 1988, 21. Gauthier joins the Humean tradition concerning the role of desires and reason in action, see page 30. Gauthier shares Mackie's metaethical view on the epistemological status of moral concepts. Hence, the concept of value is strictly bound to individual preference. Values are subjective and relative in the sense that desire determines the value of objects, e.g., their goodness and badness. Consequently, the substantive aims of an agent do not fall under the scope of rationality, they form the given goal for the achievement of which rational means must be found. Gauthier 1988, 55–59; Mackie 1990, 27.

[6] "Practical rationality in the most general sense is identified with maximization. Problems of rational choice are thus of a well-known mathematical type; one seeks to maximize some quantity subject to some constraint. The quantity to be maximized must be associated with preference; we have spoken loosely of advantage, or benefit, or satisfaction, but the theory of rational choice defines a precise measure of preference, *utility*, and identifies rationality with the maximization of utility. Utility is thus ascribed to states of affairs considered as objects of preference relations. The constraint under which utility is to be maximized is set by the possibilities of action. The rational actor maximizes her utility in choosing from a finite set of actions, which take as possible outcomes the members of a finite set of states of affairs." Gauthier 1988, 22. See also Gauthier 1988, 28. Gauthier joins the tradition according to which preferences are defined by linking them with states in the world, and not with states of mind. For a full discussion of the two alternatives, see Griffin 1990, 7–20.

ent of other people's help. The agent determined by such conditions is a *natural man* whose source of action is the set of (considered) individual preferences, and who uses rational capacity to choose the best available means for maximizing them.[9] The decision she makes in selecting her means is called *parametric choice*. She acts rationally when she chooses the best means of achieving her given ends established by her preferences.

People do not usually perform their actions and follow their pursuits solitarily but in interaction with others. The model for parametric choice does not cover such situations and for this reason the theory of practical rationality must be widened to include cases of *strategic choice*. In an interactive situation, rational individual choice emerges from common reasoning and from beliefs concerning this reasoning.[10] Interaction complicates the process of rational choice because what others do and how others reason often has unpredictable consequences for one's own prospects and opportunities. To simplify the situation, Gauthier presupposes that these interacting agents are *rational*. This implies first, that

---

[7] Gauthier stresses that his theory does not set any substantive requirements for acceptable preferences, but that preferences are individual, and that all value is, accordingly, a contingent measure of individual preference. GAUTHIER 1988, 24–25, 33. Gauthier does not even presuppose that preferences are temporally neutral; they must simply be acted on as they occur at the time of action and decision-making. GAUTHIER 1988, 37–38. Instead, Gauthier attemptsaot develop an ordinal measure for evaluating the formal acceptability of preferences. In this connection he lists four criteria for the rationality of preferences, which he calls completeness, transitivity, monotonicity and continuity. For a full explication of these, see GAUTHIER 1988, 39–45.

[8] GAUTHIER 1988, 37. See also GAUTHIER 1988, 25.

[9] Gauthier defends his individualistically defined concept of a natural man against possible dissension by claiming that irrespectively of whether we can think of a natural man independent of her necessarily social origins, it is not her origins that count but her motivations and values. We do not have to predicate complete social detachment to the natural man, but simply presuppose non-tuism: "the non-tuist takes no interest in the interests of those with whom he interacts. His utility function, measuring his preferences, is strictly independent of the utility functions of those whom he affects." This does not, however, mean that the natural man is completely indifferent towards other people, she has *socially defined secondary motives*. This means that although natural man is characterized with a vocabulary taken from social life, it is still true that: "In so far as natural man interacts with her fellows, she will exhibit certain social dispositions. But if these dispositions are based strictly on her secondary motives, then they are compatible with her underlying asociality." GAUTHIER 1988, 310–311.

each person's choice is a rational response to the choices she expects others to make, second, that she knows that others rely on similar expectations, and third, that each person believes her choice and expectations to be reflected in the expectations of every other person.[11]

A choice is rational, i.e., the conditions of strategic rationality are satisfied when two requirements, *equilibrium* and *optimality*, are fulfilled. An expected outcome is in equilibrium, if and only if it is the product of mutually utility-maximizing strategies.[12] An expected outcome is Pareto-optimal when it is not possible to improve someone's situation without, at the same time, worsening someone else's position.[13]

Gauthier defines practical rationality as maximization of expected utility. The environment in which this type of reasoning can be practiced is the *perfect market*. It is an ideal for interaction free from all constraint, an unrestrained and natural competition in which each individual pursues her own interest in her own way; she acts according to her individual preferences and relies solely on her own capacities.[14] Gauthier has borrowed the concept from Adam Smith, and along with Smith he char-

---

[10] "Rational actors must determine their choices, not in fixed circumstances, but in terms of reciprocal expectations about those very choices. Both choices and expectations must rest on the actors' beliefs about the actions each may perform, the possible outcomes of these actions, and the utilities to each of these possible outcomes. [...] In effect individual choice must emerge from common reasoning. Each must view strategic choice both as a response to the choices of his fellows and as being responded to by those choices." GAUTHIER 1988, 60.

[11] GAUTHIER 1988, 61.

[12] GAUTHIER 1988, 65.

[13] GAUTHIER 1988, 76. Gauthier admits that both conditions involve severe problems, and that neither of them can be a sufficient condition for rational choice. Gauthier is, however, convinced that part of these complications can be overcome with the help of his own moral theory. GAUTHIER 1988, 75, 77.

[14] "The perfectly competitive market presupposes private ownership of all products and factors of production. Thus the market is specified, not only by the utility functions of those interacting in it, which determine demand, and the production functions reflecting existing technology, which determine supply, but also by an initial distribution of factors, affording each person his initial factor endowment. For our purposes each person may be defined by his utility function and his factor endowment. These fix his preferences and capacities, which alone are relevant to his activity in the market." GAUTHIER 1988, 86.

acterizes the mechanism working in such unconstrained interaction as the *invisible hand*. The working of the invisible hand ensures that "the divergent and seemingly opposed interests of different individuals fully harmonize".[15] Thus, the perfect market is the opposite of the situation known as the Prisoner's dilemma: the perfect market guarantees the coincidence of equilibrium and optimality, it prevents both parasitism and free-ridership.[16]

The perfectly competitive market represents the concept of idealized interaction in which choices are made under certainty.[17] Perfect circumstances make restrictions of individual utility-maximizing choice unnecessary: free activity under certainty ensures that the market constantly moves to an equlibrium and to an optimal state.[18] Under the conditions of a perfectly competitive market there are no conflicts between mutual benefit and the pursuit of individual gain for the perfect market realizes both in parallel. Consequently, the free market ensures every individual, at the interpersonal level, the same amount of freedom that Robinson Crusoe enjoyed in solitude. In addition, it secures them the vast benefits of economic interaction and exchange.[19] The market mechanism naturally guarantees an optimal outcome in co-operative interaction, and for this reason, it represents *a morally free zone*. The market is non-moral because people's activities are compatible with each other. It forms an area of freedom rationalized by the optimality of its functioning.[20] Gau-

---

[15] GAUTHIER 1988, 83.

[16] "Conceived as an ideal type, the perfect market, [...], guarantees the coincidence of equilibrium and optimality, and so its structure is the very antithesis of the Prisoner's Dilemma." GAUTHIER 1988, 83. See also GAUTHIER 1988, 96.

[17] "The idealization of interaction represented by the perfectly competitive market includes the removal of both circumstantial uncertainty and strategic calculation. Market decision-making, although practically complex, is logically of the simplest kind. Each chooses parametrically, as if his actions were the sole variable factor, taking the actions of others as fixed circumstances. And each chooses as if he knew the outcome of each of his possible actions. The core of traditional economic theory, examining the workings of the market, is thus concerned with the analysis of rational choice under the assumption of certainty." GAUTHIER 1988, 85.

[18] GAUTHIER 1988, 84, 89–90.

[19] GAUTHIER 1988, 90, 93.

thier's idea of the market as a morally free zone reflects his conception of morality as a constraint, or a restriction. Morality is not a scheme of the good life to be adopted for realizing value that non-moral, or immoral life lacks, but a restriction imposed on the natural human being for the purpose of gaining some preferred goal.

Summing up, the ideal situation for human action is the one in which an agent succeeds in maximizing her utilities determined by her preferences without being inhibited by her psychological conditions, by other agents, or by restrictive rules. From this perspective, self-interest is the driving and directing force of intentional human action, or the key model of rationality. Here an agent whose aspirations correspond with the conditions of this ideal situation is a *natural individual*, whose identity is based on her preferences and on her plans of action for realizing them. This is the basic concept of a person in Gauthier's theory and his moral theory is designed for such persons.

2.2.2. The rationality of co-operation or a model for an *economic man*

The perfect market is an idealized, not a realistic picture of rational interaction between human beings.[21] In reality, there are factors which distort

[20] Gauthier 1988, 103. Gauthier lists the characteristics of an individual within the perfect market: first, the market ensures her free activity in the sense that she is as free to act as if she were as a solitary individual. Second, her condition is not influenced by externalities, which means that she is not affected by any market activity to which she has not chosen to be party. The optimality of the market secures the individual at least as much benefit as she would have as a solitary being. These three conditions together guarantee that the market is impartial. Furthermore, the parties are mutually indifferent to each other. Mutual unconcern has two effects: first, it secures the agent a fundamental liberty, equal to the liberty of Robinson Crusoe; second, it facilitates a new, wider perspective on mutual co-operation by ignoring the conventional ties of affection and friendship. These conditions together establish what Gauthier calls a person's *market self*. Gauthier 1988, 96–101.

the functioning of the perfect market. The need for ethics in the form of moral constraints arises from this market failure.[22] Factors which distort the perfect market are called *circumstances of justice*. They give rise to both co-operation and mutual concern, i.e., to the virtue of justice. A new type of interaction becomes necessary because external circumstances do not fulfil the criteria of the perfect market, and because people are biased towards themselves.[23] But the situation can be corrected with a mechanism that re-establishes the conditions of the perfect market. People acknowledge the beneficence of co-operative interaction, but they recognize that co-operation can also be used for exploiting others. For this reason, there must be a constraint which secures the advantages of co-operation but inhibits exploitation, i.e., there must be justice. Accordingly, justice is "the disposition not to take advantage of one's fellows, not to seek free goods or to impose uncompensated costs, provided that one supposes others [to be] similarly disposed."[24]

The conditions of justice place practical rationality in a new light because the circumstances which determine rationality for market interaction, are deficient for interactive situations. It is impossible to define a rational course of action if one cannot anticipate the outcome of the action. Such anticipation becomes unfeasible unless there are agreed conditions of interaction, which enable the parties to predict how others will react to what they do. When an agreement on this has been reached, a new mode of interaction, *co-operation*, has evolved.[25]Furthermore, co-operation is rational if it realizes the same conditions as the ideal market determined by perfect competition. The market mechanism, which

---

[21] Peculiarly, Gauthier points out this condition as idealized, but ignores other equally non-realistic features of his presuppositions, e.g., he attributes rationality to agents without problematizing this assumption.

[22] GAUTHIER 1988, 84.

[23] "[...] we may then say that the fundamental circumstances of justice, those features of the human situation that give rise to co-operation, are awareness of externalities in our environment, and awareness of self-bias in our character." GAUTHIER 1988, 116. 'Externalities' are "unchosen third-party costs and benefits". GAUTHIER 1988, 116.

[24] GAUTHIER 1988, 113, 116.

[25] GAUTHIER 1988, 117.

secures optimality and equilibrium, represents the invisible hand, whereas factors determining rational co-operation function as the *visible hand*; both have the same effect on individual action and decision-making.[26]

Rationality of co-operation consists in two components. *Internal rationality of co-operation* defines the conditions under which the agreement concerning co-operation is rational. *External rationality of co-operation* determines when it is rational to agree to act co-operatively rather than non-co-operatively. External rationality is, thus, a presupposition for internal rationality: one must first evaluate whether co-operative action is more rational than a solitary activity, and if it is, determine the most rational form of co-operation. When these two conditions of rationality have been settled it can be decided whether the model of co-operation based on an agreement is *moral*. This issue occupies a central place in Gauthier's theory for it involves a move from rationality to morality.[27]I examine Gauthier's solution to these three concerns.

A solitary agent acts rationally, or fulfils the criteria of internal rationality, as long as her action is utility-maximizing. The same must somehow also apply to co-operational activity: co-operation cannot be rational unless it reflects the preferences of each person.[28] The main difference between solitary and co-operative situations is that in interactive cases the conditions of rational choice also depend on the structure of interac-

---

[26] "Co-operation arises from the failure of market interaction to bring about an optimal outcome because of the presence of externalities. We may then think of co-operative interaction as a visible hand which supplants the invisible hand, in order to realize the same ideal as the market provides under conditions of perfect competition." GAUTHIER 1988, 128.

[27] GAUTHIER 1988, 118–119.

[28] "Several persons are to agree on an outcome which is then to be brought about by a joint strategy determining each person's action. Under what conditions is their agreement rational? Individual choice is rational in so far as it is utility-maximizing. Is there an analogue to utility-maximization for agreement or co-operative choice? Since it is to be voluntary, it must reflect, and in some sense reflect equally, the preferences of each person. A first proposal, therefore, would be that we derive, from the individual preference orderings of the co-operators, a social preference ordering. We then define a measure of social preference, and identify rational co-operative choice with the maximization of this measure." GAUTHIER 1988, 122–123.

tion, whereas a solitary agent can concentrate solely on the effects of her own action.[29]

Rational agents, who make decisions in interpersonal situations, purport to adopt that joint strategy selected for co-operative interaction which will secure them as many benefits and as few burdens from the co-operation as possible. This being everyone's aim, rational co-operation involves curbing one's liberty so as not to take undue advantage of others, and in return, avoiding being taken advantage of by others. Consequently, the search for a rational co-operative strategy is a *process of bargaining*, the outcome of which secures both the rationality and the optimality of co-operation.[30]

The procedure for bargaining consists of two principal stages which follow each other in turn until the final agreement is reached or the parties refrain from making an agreement. First, each party makes a suggestion concerning the final outcome or agreement. If, and this is usually the case, the claims are incompatible, the process proceeds to the second stage in which at least one party offers a concession by modifying the initial claim and proposing an alternative outcome.[31]

The rationality of possible claims and concessions is determined by the initial assumptions of the bargaining parties: on the basis of individual rationality each wants as much as possible. To get as much as possible presupposes claiming as much as possible, while the outcome is determined by the claims forwarded and concessions made. The claims cannot, however, be unrestricted, for advancing too demanding a claim endangers the whole process of bargaining: no one will want to take part if the outcome will not secure her at least the amount of goods she would attain were she not involved in co-operative interaction. Consequently, one's claim is restrained by the extent of her participation in co-

---

[29] "Co-operative choice must reflect the preferences of each person, if it is to be rational, but how it reflects their preferences must depend on the structure of their interaction, on the consequences for everyone of what each is able to do." GAUTHIER 1988, 126.

[30] GAUTHIER 1988, 128.

[31] GAUTHIER 1988 133.

operative interaction: the benefits she can reasonably expect must be relative to the costs she has suffered in realizing the overall co-operative surplus.[32] These conditions establish the internal rationality of co-operation. From this point of view, co-operation is a means for the utility-maximizing agent to achieve her goals, and she will accept an agreement on co-operation, and act accordingly, if this is a better means for her to get what she wants than a solitary pursuit.[33]

The conditions determining the internal rationality of the bargaining position also mark the conditions of its external rationality. The principle which fulfils the criteria of external rationality is *minimax relative concession*; it determines the best outcome of a bargaining procedure. The idea behind this principle is that the concessions of different bargainers are measured by using the concept of relative concession, comparing the magnitude of the concessions each party makes in relation to their initial suggestions in terms of utility-differences.[34] The advantage of this solution is that it makes the theoretically problematic interpersonal comparisons unnecessary.[35] The minimax relative concession principle states that for any co-operative interaction we must choose the outcome in which the maximum relative concession required from the different parties is as small as possible.[36] If this condition is fulfilled, it is rational to join in the co-operation.[37]

---

[32] GAUTHIER 1988, 133–134.

[33] Interestingly, Gauthier describes the process which eventually leads to an agreements as bargaining, whereas in Rawls depicts it as a negotiation.

[34] "However, we may introduce a measure of *relative concession* which does enable us to compare the concessions of different bargainers, and which thus gives us a basis for determining what concession each must rationally make." GAUTHIER 1988, 134–135. "Relative concession is a proportion of two utility-differences or intervals. [...] Relative concession is independent of the choice of utility scale. Each person's relative concessions are fixed no matter how we choose to measure his utilities." GAUTHIER 1988, 136.

[35] "Given the claims of the bargainers, what concessions is it rational for them to make? To answer this question we must first consider how concessions are to be measured. The absolute magnitude of a concession, in terms of utility, is of course the difference between the utility one would expect from the outcome initially claimed and the utility one would expect from the outcome proposed as a concession. But this magnitude offers no basis for relating the concessions of different bargainers, since the measure of individual utility does not permit interpersonal comparisons." GAUTHIER 1988, 134.

The conditions which determine the external and internal rationality of a co-operational agreement also draw the portrait of an *economic man*. Economic man co-operates on a rational basis, but the interaction with others is asocial; she associates with others for purely individual reasons and treats co-operation only as a means for achieving her own ends. Economic man accepts a moral constraint to restrict her activity because a mutual moral agreement guarantees her a better outcome in her utility-maximization than a contractless state, but her relation to morality is as instrumental as her contact to other people. She will comply with morality, but her utility-maximizing concept of rationality tells her to break with morality when confiding in it would cause her disadvantage.[38] From the point of view of an economic man morality does not offer any independent premises for practical rationality or any specific reasons for action. It is simply a part of utility-maximizing rationality for interactive situations.

2.2.3. MORALITY AS CONSTRAINED MAXIMIZATION — A *LIBERAL INDIVIDUAL* EMERGES

The principle of minimax relative concession also plays a third role in Gauthier's theory: it represents the principle of rational behaviour in co-operative interaction. The minimax relative concession principle entails a transition from individual utility-maximizing rationality to a new type of rationality.[39] This is the move from rationality to morality. Gauthier pur-

[36] "[...] we claim that the principle should state that given a range of outcomes, each of which requires concessions by some or all persons if it is to be selected, then an outcome be selected only if the greatest or *maximum* relative concession it requires is as small as possible, or a *minimum*, that is, is no greater than the maximum relative concession required by every other outcome. We call this the principle of miminum-maximum, or *minimax relative concession*." GAUTHIER 1988, 137.

[37] GAUTHIER 1988, 145.

[38] GAUTHIER 1988 316–324.

ports to show that it is rational (according to the criteria of internal and external rationality) to adopt minimax relative concession as one's action guiding principle, but that accepting this principle entails, additionally, a change in one's concept of rationality. This means grounding morality on rational utility-maximizing considerations and demonstrating thereby that it is rational to be moral even independently of utility-maximization.[40] In other words, to show that the minimax relative concession principle is a basis for rationality in co-operative interaction is to prove that, in terms of utility-maximization, it is rational to be moral. In the ideal of co-operative interaction constrained by the minimax relative concession principle justice and rationality coincide.[41] This is the heart of Gauthier's theory, for him the whole project of a reason-based morality depends on the possibility of connecting practical reason and morality through utility-maximization.[42] But this is also the point where the problematic nature of Gauthier's theory becomes evident. The economic

[39] "[...] the principle of minimax relative concession is the principle of rational behaviour in co-operative interaction — interaction based on the joint strategy agreed to in bargaining. Each person acts, not to maximize his own utility, but to bring about the outcome that is the object of the bargain, affording each person an expected utility no less than he would expect from his maximal claim and minimax concession." GAUTHIER 1988, 145.

[40] "It is this third role which establishes the distinctively moral character of the principle, and of co-operation. For applied to co-operative interaction, the principle of minimax relative concession constitutes a constraint on the direct pursuit of individual utility. Thus if we can show it to be a rational and impartial basis for co-operative interaction, we shall have established its credentials as a moral principle." GAUTHIER 1988, 145.

[41] "[...] justice and reason coincide in a single ideal of co-operative interaction. The principle of minimax relative concession serves not only as the basis for rational agreement, but also as the ground of an impartial constraint on each person's behaviour. And justice is the disposition to abide by this constraint. In treating co-operative interaction as the domain of justice we make a twofold claim. The first is that like the market, rational co-operation excludes all partiality. The second is that unlike the market, this exclusion requires each co-operator to constrain her maximizing activity." GAUTHIER 1988, 150. See also GAUTHIER 1988, 158.

[42] "If our defence fails, then we must conclude that rational bargaining is in vain and that co-operation, although on a rationally agreed basis, is not itself rationally required, so that it does not enable us to overcome the failings of natural and market interaction. Indeed, if our defence fails, then we must conclude that a rational morality is a chimera, so that there is no rational and impartial constraint on the pursuit of individual utility." GAUTHIER 1988, 158.

man should now be convinced that it is rational to join in co-operation and to comply to rules which regulate this co-operation under certain conditions. The next step is to show that it is rational to keep to these same rules even in situations in which compliance does not necessarily produce the best utility-maximizing result. The problem is that the economic man will only be convinced of the usefulness of obeying the rules if this can be shown in terms of utility-maximizing. But understanding the real nature of these rules (for they are after all moral rules) means that one does not regard them as an instrument for one's own benefit, but as something that deserves compliance in its own right. Gauthier cannot find a satisfactory solution to this dilemma in his theory.

To show that one can benefit from keeping agreements when it suits oneself is not a sufficient condition for the rationality of morality. There has to be a rational basis for being *morally disposed* towards constraints, which regulate co-operative interaction. This means showing that it is rational, not just to use morality as a means for utility-maximization, but to be moral because of morality itself.[43]

Gauthier calls his model for rational co-operation *constrained maximization*. A constrained maximizer maximizes her utility given the inner moral restrictions which direct her action towards other people. Gauthier defends constrained maximization against an egoist's straightforward maximization, which involves applying individual utility-maximizing rationality directly to interaction.[44] A straightforward maximizer, the embodiment of which is the economic man, is a person who yields to agreements only when this squares with her own interests.[45] Straightforward maximization does not, however, guarantee that a chosen strategy is rational in a situation of co-operation, because the presence of others

---

[43] GAUTHIER 1988, 165.

[44] In his argumentation Gauthier follows Hobbes: the objection against the rationality of morality does not arise from the source of suspecting the usefulness of agreements; all utility-maximizing agents acknowledge the beneficial nature of contracts. What we must show is that there is a reason for *being* moral and not just acting morally when it produces the greatest benefit in pursuit of one's greatest utility. GAUTHIER 1988, 160–161.

[45] "Let us say that a *straightforward* maximizer is a person who seeks to maximize his utility given the strategies of those with whom he interacts." GAUTHIER 1988, 167.

complicates the situation. First, the straightforward strategy endangers co-operation, because it cannot secure a joint strategy necessary for this form of interaction.[46] Second, straightforward maximization is not a rational strategy for co-operative interaction because only agents disposed to keep their agreements are rationally acceptable as parties to agreements. Straightforward maximizers do not fulfil this condition, for they are not morally disposed and are, thus, known to be liable to violating agreements. Such persons will not be accepted as parties to beneficial co-operative ventures based upon mutual agreement.[47]

Constrained, not straightforward maximization, is the rational strategy for co-operative action. A constrained maximizer builds her co-operative strategy on the maximization of other parties' utilities, not her own. She will be morally disposed, and act accordingly, irrespective of any individual strategy which could produce her greater expected utility.[48] She will not, however, let her actions be constrained if the co-oper-

---

[46] "We may think of participation in a co-operative activity [...] as the implementation of a single joint strategy. [...] An individual is not able to ensure that he acts on a joint strategy, since whether he does depends, not only on what he intends, but on what those with whom he interacts intend. [...] A person co-operates with his fellows only if he bases his actions on a joint strategy; to agree to co-operate is to agree to employ a joint rather than individual strategy. [...] A joint strategy is fully rational only if it yields an optimal outcome, or in other words, only if it affords each person who acts on it the maximum utility compatible in the situation with the utility afforded each other person who acts on the strategy. Thus we may say that a person acting on a rational joint strategy maximizes his utility, subject to the constraint set by the utilities it affords to every other person. An individual strategy is rational if and only if it maximizes one's utility given the *strategies* adopted by the other persons; a joint strategy is rational only if (but not if and only if) it maximizes one's utility given the *utilities* afforded to the other persons." GAUTHIER 1988, 166–167.

[47] In discussing the argument in favour of straightforward maximization as a rational utility-maximizing strategy for co-operation, Gauthier maintains: "[...] this argument would be valid only if the probability of others acting co-operatively were, as the argument assumes, independent of one's own disposition. And this is not the case. Since persons disposed to co-operation only act co-operatively with those whom they suppose to be similarly disposed, a straightforward maximizer does not have the opportunities to benefit which present themselves to the constrained maximizer. [...] [Straightforward maximizers] would not be admitted as parties to agreement given their disposition to violation. Straightforward maximizers are disposed to take advantage of their fellows should the opportunity arise; knowing this, their fellows would prevent such opportunity arising." GAUTHIER 1988, 172–173.

ative situation is likely to be unfair for her. A rational agent is only *conditionally disposed* to act morally, and she will move from constrained maximization to strategies of individual maximization if there is a strong likelihood that others do not co-operate. An agent must compare the expected utility produced by the estimated degree of co-operation from others with the expected utility of universal non-co-operation. If the expected utility of universal non-co-operation is greater a rational agent will adopt an individual strategy and ignore her moral disposition to constrain utility-maximization.[49] This does not mean abandoning morality, but choosing a course of action in which moral questions do not rise.

According to Gauthier, constrained maximization is not just a more efficient version of straightforward maximization.[50] In that case, morality would simply be a means for realizing non-moral ends. Constrained maximization is rational but not for the reason that it would be a more effective instrument for realizing one's utility than straightforward maximization; the real difference between these two is the *variety of choice* which the agent's disposition secures her. The argument for straightforward maximization ignores the effect of disposition on one's variety of choices. Thus, a constrained maximizer benefits from her disposition, because it opens up new opportunities for her.[51] To explicate, straightforward maximization represents purely instrumental rationality, whereas constrained maximization is a wider scheme of rationality comprising the rationality of ends in the sense that it increases the variety of alternatives open to the agent. Constrained maximization connects util-

---

[48] "A *constrained* maximizer [...] is a person who seeks in some situation to maximize her utility, given not the strategies but the utilities of those with whom she interacts. [...] A constrained maximizer has a conditional disposition to base her actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility." GAUTHIER 1988, 167.

[49] GAUTHIER 1988, 167–168.

[50] "[...] constrained maximization is not straightforward maximization in its most effective disguise. The constrained maximizer is not merely the person who, taking a larger view than her fellows, serves her overall interests by sacrificing the immediate benefits of ignoring joint strategies and violating co-operative arrangements in order to obtain the long-run benefits of being trusted by others. Such a person exhibits no real constraint." GAUTHIER 1988, 169–170. See also GAUTHIER 1988, 188–189.

ity-maximization, as the core of practical rationality, to morality. This forms a basis for a new definition of rationality as utility-maximization at the level of dispositions to choose.[52] Being morally constrained does not necessarily secure a greater amount of utility, but it produces a wider and novel variety of utilities.

Adopting a new concept of rationality also involves a change in the concept of a person. Straightforward maximization is the form of rationality adopted by the economic man, but constrained maximization as an action-guiding principle presupposes a different concept of a person, that of a *liberal individual*. Constrained maximization involves a choice to keep the rational contract and to become disposed in a certain way, moving from instrumental motivation to genuine sociality. The choice covers both the form of one's rationality and one's self-interpretation. It is a transformation from an economic man to a liberal individual.[53] Before we can examine the liberal individual more closely, it is, nevertheless, necessary to study Gauthier's idea for constructing the agreement in more detail.

---

[51] "It might seem that a maximizing disposition to choose would express itself in maximizing choices. But we have shown that this is not so. The essential point in our argument is that one's disposition to choose affects the situations in which one may expect to find oneself. A straightforward maximizer, who is disposed to make maximizing choices, must expect to be excluded from co-operative arrangements which he would find advantageous. A constrained maximizer may expect to be included in such arrangements. She benefits from her disposition, not in the choices she makes, but in her opportunities to choose." GAUTHIER 1988, 183.

[52] "A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition." GAUTHIER 1988, 182–183.

[53] GAUTHIER 1988, 328–329.

2.2.4. JUSTIFYING THE CONTRACT

Practical rationality in an interactive situation is based on co-operation. A joint strategy enabling such co-operation is, as we have seen, an outcome of a bargaining process in which offers and concessions follow each other. Now, how is an agreement on morality reached?

Theories of the social contract tradition usually contain a concept of the conditions which precede the moral agreement, a *state of nature*. This is, however, only an imaginary state; the existence of an ordered society and the possibility of mutual co-operation show that people have a common basis which they accept as limiting their natural freedom. The rationale which sustains society and morality can be made understandable by contrasting society with a state of nature and by constructing the concept of a *proviso* for defining the condition under which it is rational to make an agreement.[54]

The most important feature of the proviso is that it guarantees non-coerciveness, for it is rational to comply with a bargain only if the compliance is reached without duress.[55] For this reason, the proviso must guarantee that no one can take advantage of others.[56] This means, Gauthier maintains, that the proviso must guarantee each person the "ownership" of her own body and its powers and capacities, that is, her *basic endowments*.[57] Basic endowments form a base utility, which is not included in the co-operative surplus, and which is initially acquired without taking advantage of any other person.[58] Accordingly, each individual may expect to use her own powers and capacities for her own good, but in

---

[54] "We shall therefore argue that it is both rational and just for each individual to accept a certain constraint on natural interaction, and on the determination of his initial factor endowment, as a condition of being voluntarily acceptable to his fellows as a party to co-operative and market arrangements — to social interaction. This constraint is part of morals by agreement, not in being the object of an agreement among rational individuals, but in being a precondition to such agreement." GAUTHIER 1988, 192. "We may say that the proviso moralizes and rationalizes the state of nature — but only in so far as we conceive the state of nature as giving way to society. [...] Without the prospect of agreement and society, there would be no morality, and the proviso would have no rationale. Fortunately, the prospect of society is realized for us; our concern is then to understand the rationale of the morality that sustains it." GAUTHIER 1988, 192–193.

doing this she may not expect to use other persons' powers to her advantage. This basic utility moulds the unlimited liberties of the initial position into exclusive rights and duties: each person has an exclusive right to use her own powers without being prevented by others, and a corresponding duty to abstain from using other persons' powers in a manner inhibiting the use of their powers for their own good.[59] These basic endowments define the basic rights upon whose foundation it is rational to make an agreement. They also form, what Gauthier calls, a person's *normative sense of self*. This sense is normative for Gauthier maintains that each person identifies with her physical and mental capacities to which she has direct access.[60] Interestingly, the normative sense of the self includes the same elements as the natural man, discussed above.[61] They both define, for this theory, the conditions necessary for intentional action.

Gauthier defines the scope of the proviso further by maintaining that

[55] Gauthier exemplifies this condition with an imaginative story of a contract made between masters and their slaves: the deal is to secure a better and more willing service for the masters by the slaves and improved living conditions and more freedom for the slaves. After the contract has been made, the masters who had suggested the agreement realize that the arrangement makes everything worse; instead of being content with their new condition, the slaves now demand much further-going freedoms and rights for themselves. The rationale of the story is this: the agreement the masters made with the slaves was not rational because the initial bargaining situation was coercive. In the story, any bargain ensuring the slaves' improvement was rational from their point of view during the conditions of slavery, whereas in the new situation they can aim — and it is rational for them to aim — at fuller liberties and possibilities to influence their own life-circumstances. A rational agreement is stable, and it cannot be stable unless it is made under non-coercive initial conditions. See GAUTHIER 1988, 190–192.

[56] GAUTHIER 1988, 192.

[57] GAUTHIER 1988, 208.

[58] GAUTHIER 1988, 200–201.

[59] "Each person, in the absence of his fellows, may expect to use his own powers but not theirs. This difference is crucial. For it provides the base point against which the proviso may be applied to interaction. [...] Thus the proviso, in prohibiting each from bettering his situation by worsening that of others, but otherwise leaving each free to do as he pleases, not only confirms each in the use of his own powers, but in denying to others the use of those powers, affords to each the exclusive use of his own. [...] Each person has an exclusive right to the exercise of his own powers without hindrance from others, and a duty to refrain from the use of others' powers in so far as this would hinder their exercise by those with direct access to them." GAUTHIER 1988, 209–210.

although an agent has right *to* her powers she has a right only *in* the effects of her labour, i.e., not an exclusive right to their possession. This is due to the basic conditions of the proviso: one may improve one's own situation provided that one does not thereby worsen anyone else's situation.[62] This reasoning develops the initial rights to include rights of property. Gauthier justifies the right to private property by referring to the beneficial effects possessive rights have to communal good.[63]

To summarize: the imaginary initial position, which precedes society, is determined by the proviso which generates a set of rights for each person: a right to her person and in the fruits of her labour [sic], and a right to personal property. In keeping with these rights, the proviso also determines the rationality of the parties so that they accept the criteria for the internal and external rationality of co-operation.[64] The proviso is further characterized by the agents' aim to maximize their utilities, along the

---

[60] GAUTHIER 1988, 210. Gauthier stresses that, in the contractarian tradition, rights are not an outcome of the agreement but they form the starting point for it. GAUTHIER 1988, 222. Gauthier dismisses the view that natural distribution is a fair starting point for an agreement: "the initial bargaining position, as the starting point for rational co-operation, may not be identified with the natural distribution, or non-co-operative outcome, the natural distribution represents the effects of power. [...] If we think of each person's endowment as constituting her rights [...] then we may say that to accept natural distribution as the initial bargaining position would be to accept might as making right." GAUTHIER 1988, 198–199.

[61] See page 130.

[62] One may not seize the products of another person's labour, without providing a full compansation for the producer, while in doing so the other person deprives the producer not only of the product of her labour but also of her intended use of the good. On the other hand,one is free to use the product deriving from another person's work which the producer has not laboured to produce. Moreover, one is also free to take advantage of products of somebody else's labour when the producer herself could find no use whatsoever for them. GAUTHIER 1988, 210–211.

[63] GAUTHIER 1988, 216–217.

[64] "[...] interaction constrained by the proviso generates a set of rights for each person, which he brings to the bargaining table of society as his initial endowment. He brings a right to his person, a right in the fruits of his labour [sic], and a right to those goods, whose exclusive individual possession is mutually beneficial, that he has acquired either initially or through exchange. [...] Without these rights, persons would not be rationally disposed, either to accept the prohibition on force and fraud needed for market competition, or to comply voluntarily with the joint strategies and practices needed for co-operation." GAUTHIER 1988, 227. See also GAUTHIER 1988, 222–223.

requirements of minimax relative concession. The parties' utility-maximization is not straightforward, for co-operation in the initial position does not take place under conditions of perfect competition.[65] Last, the agents' disposition to constrain their action morally is only conditional: they are not required, in terms of rationality, to act morally when it is likely that others do not constrain their action similarly.[66]

To interpret Gauthier's concept of the proviso, we can say that it is the point at which rationality and morality coincide. In the proviso, rational concern changes into moral concern. Understanding oneself as a rational party who enters a contract, means changing both one's basic disposition toward co-operation and one's concept of rationality. The fact that the proviso is not itself an object of the contract but its prerequisite shows that the moral perspective, necessary for rational co-operation, is something that must be chosen and adopted. Instead of conceiving oneself purely in terms of economic, or straightforward maximizing rationality, an agent must adopt a different perspective on herself and on her rationality. Although it is rational, even from the straightforward maximizing perspective, to enter the proviso, the outcome produced by the proviso is a new mode of rationality which cannot fully be estimated by the criteria of economic rationality. We can also say that accepting the conditions of the proviso as determining human rationality and action, means realizing that interactional situations change the con-

---

[65] Gauthier 1988, 223.

[66] Gauthier distinguishes two different levels of compliance, broad and narrow, to examine how far it is rational to comply with the moral constraints given the expected degree of compliance in others. Gauthier defines broad compliance by saying that: "[a] person disposed to broad compliance compares the benefit she would expect from co-operation on whatever terms are offered with what she would expect from non-co-operation, and complies if the former is greater." Gauthier 1988, 225–226. A disposition corresponding with narrow compliance, again, involves that a person "compares the benefit he would expect from co-operation with what he would expect from a fair and optimal outcome, and complies with a joint strategy only if the former approaches the latter." Gauthier 1988, 226. A rational agent should be only narrowly compliant: "It is rational for each person to be sufficiently compliant that society is possible if others are equally compliant; it is not rational for anyone to be so compliant that society is possible if others are less compliant; therefore it is rational for each person to be narrowly compliant." Gauthier 1988, 227.

ditions of intentional action. Straightforward rationality or rationality of means is not enough for determining the best means to achieve a desired end; a new mode of consideration is needed. Intentional action becomes specifically moral action.

2.2.5. The Archimedean point of morality

We have seen, says Gauthier, that the proviso is a rational basis for co-operation, and that it connects morality with rationality. The proviso does not, however, constitute a moral theory. In addition, there has to be an *Archimedean point* for governing the moral realm.[67] The Archimedean point is characterized through its occupant who is an ideal rational actor. The task is to show that this ideal actor would accept the proviso as the initial condition of constraint and constrained maximization as the strategy for defining the basic structure of society. This means that the concept of an ideal actor must secure both impartiality of moral decisions and an outcome of choices with which each actual person could identify herself.[68] The Archimedean point guarantees that the notion of "morals by agreement" is universalizable and individually acceptable.

The choices made in the Archimedean point are impartial because its occupant, the ideal actor, is unaware of her particular identity, and thus, of her particular conception of the good. This does not, however, mean that the ideal actor lacks individuality: she knows that she is an individual with her own view of the good, with her own capacities and aspirations; what she does not know is the specific content of this individuality.[69] Gauthier stresses this point, while the knowledge of individuality is a necessary condition of choice: one has to be someone in order to be able

[67] Gauthier 1988, 233. Gauthier borrows this concept, as so many of his central ideas, from Rawls, see Rawls 1972, 260–263, 584.

[68] Gauthier 1988, 235–236, 259–260.

[69] Gauthier 1988, 234, 244.

to choose something.[70]

An ideal agent is an individual and this makes her concerned with her identity whatever that identity may be. Her choice of principles for social interaction reflects this concern. This unspecified individuality notwithstanding, she cannot tailor principles to suit her particular condition, but she can choose principles which any rational agent would choose to further her own good, whatever that might be.[71] Gauthier's view implies that the ideal actor will not make her choices as if she had an equal chance of being each of the persons affected by the choice, but as if she were each of those persons. Her choice represents a point of convergence, in which everyone can identify with the choice even though particular individuality loses its significance.[72] In Gauthier's theory this "non-biased" individuality has a decisive impact on the choice of principles of social justice. To protect her individuality, whatever its particular content, the ideal actor does not, unlike the Rawlsian ideal actor, choose the *lexical difference principle* to regulate the distribution of societal goods, but the principle of minimax relative concession.[73]

The ideal actor reasons from the conditions mutual to all individuals. Consequently, if the ideal actor chooses to benefit herself, then she must choose mutual benefit; if freedom, then freedom for everyone.[74] The initial conditions of rationality in the form of constrained utility-maximization and ignorance concerning her particular identity as her premises, the ideal actor accepts the proviso as one of the principles for interaction.[75]

---

[70] Gauthier criticizes Rawls for his aim to nullify the effect of contingencies of nature on principles of justice. Depriving the parties in the original position of all contingency, also robs them of their individuality; human individuality simply cannot be separated from contingencies. GAUTHIER 1988, 237. The individuality of the ideal actor must, furthermore, be preserved for the concept of choice to make sense: "When the ideal actor chooses among feasible principles of interaction, she does so, let us not forget, as an actor. Her choice is that of a person with particular capacities, attitudes, and preferences, even though it is made in ignorance of what these are. But had she no preferences to order possible outcomes, no capacity to choose among possible actions, and no understanding of the relation between actions as objects of choice and outcomes as objects of preference, then she would not be an actor. She would not be implicated in the choice among principles of interaction." GAUTHIER 1988, 253–254.

[71] GAUTHIER 1988, 251.

[72] GAUTHIER 1988, 255–256.

The principles of social interaction chosen by the ideal actor in this situation are the same as the principles of rational co-operation, which are, again, the principles of constrained maximization.[76]

At bottom the Archimedean point has two central features: it guarantees both the universalizability and the acceptability of the chosen principles. Is is a warrant of universalizability because it is void of any substantive content. The concept of justice it produces is purely instrumental: it does not dictate a concept of the good, and it does not exclude any individual goals. It only requires that people accept a rationally grounded system of mutual rights and constraints.[77] The second feature of the Archimedean point, acceptability, represents Gauthier's view on the moral person: the characteristics of the ideal rational actor are the properties we, as moral persons, have when we reason morally. The Archimedean point is an instrument of moral thought in the same sense as rationality is an instrument of deliberation.[78]

---

[73] GAUTHIER 1988, 246. Rawls justifies the choice of the lexical difference principle by referring to the cruciality of the lot of those worst off. In Rawls' view all natural differences as undeserved, and thus taking part in mutually beneficial co-operation does not involve a right to the good. Consequently, no one is allowed to improve their own situation on the basis of will, talent and labour, regardless of whether it worsens anyone else's situation. See RAWLS 1972, 101–102. For Gauthier Rawls' position means "that an individual's contribution does not entitle him to any return from society. [...] If talents constitute a common asset, then no individual's talent, or his effort based on that talent, affords him any particular claim to the products of social interaction." GAUTHIER 1988, 249. This is simply a wrong view: one has an initial right to one's natural assets, and this is reflected in the concept of an ideal actor: she knows that she is an individual with particular characteristics, although she does not know what these are. GAUTHIER 1988, 251.See also page 120

[74] GAUTHIER 1988, 257.

[75] "Supposing the possibility of mutual benefit, the ideal actor must choose to prohibit the unilateral taking of advantage. No one may better himself through interaction that worsens the position of another, where the base point in relation to which we determine bettering and worsening is the absense of the other party to the interaction. And so the ideal actor must choose the proviso as one of the principles for interaction." GAUTHIER 258–259. On the basis of the proviso, a fully competitive market then emerges, as an arena of social co-operation from which both force and fraud are eliminated, and which realizes optimality through everyone paying, and paying only, the costs for the goods produced through co-operation she receives. GAUTHIER 1988, 261.

[76] GAUTHIER 1988, 265.

[77] GAUTHIER 1988, 344–345.

We can now ask, in what way or ways we are qualified by the properties of the occupant of the Archimedean point, whether we relate to them in an *indicative* manner (in the sense that we are such moral persons) or in an *imperative* mode (meaning that we should become such moral persons). The answer to these questions lies in the way Gauthier defines our relation to the proviso. Adopting the Archimedean point, or accepting the moral position, means *choosing oneself as a moral person*. Thus, although it is rational to accept the moral perspective, one still has to choose it, concede to a different mode of rationality. Despite Gauthier's assurances that a rational morality is possible, there seems to remain a gap between rationality and morality. This gap is bridged by the choice one makes in adopting the proviso. In this choice morality ceases to be a mere means and becomes an intrinsic human value. Gauthier purports to span this gap by introducing a concept of a *liberal individual*. The difference between the rational and the moral point of view corresponds with the contrast between the economic man and the liberal individual.

The liberal individual is defined basically by the same characteristics as the economic man. Consequently, her individual concept of the good and her preferences determine her goals. What distinguishes her from the economic man, however, is the way in which she sees her life as fulfilling her preferences; her striving towards her own particular ends does not just have an instrumental value for her, but her ends are intrinsically valuable to her. This colours both her attitude towards morality and towards other human beings: they represent a value in themselves. Gauthier defines the liberal individual from the perspective opened by the moral point of view, but he believes that the liberal individual depicts the real, empirical human being better than the image of an economic man.[79]

Gauthier's moral theory can be seen as an attempt to avoid the difficulties which arise from a definition of a substantive good, by introducing a purely formal moral theory. The formal theory is then made adaptable for people with their individual concepts of the good by intro-

[78] Gauthier 1988, 344.

[79] Gauthier 1988, 346–347, 353–354.

ducing the concept of an ideal rational agent. The theory remains formal, but the formal features are such that everybody can, and must if they are rational, accept them as the basis of their acting. But are the basic concepts of Gauthier's theory as void of substantive content as he presupposes? Does not the concept of the liberal individual smuggle in material that cannot be accepted on purely formal grounds? Does it not represent a view of the substantive good in human life, a view on human flourishing? It seems that we necessarily have to overcome the gap between morality and rationality by a choice of ourselves as moral persons cannot be bridged unless we accept a certain substantive view on good human life, a view on life as a liberal individual.

2.3. The rationally justifiable morality of Alan Gewirth

Gewirth[1] joins the tradition of moral philosophy, which attempts to find a rational justification for the institution of morality. As we have noticed during the course of this study, several moral theorists see it as their task to reconcile morality with rationality. These thinkers, e.g., Brandt, wish to show that being moral does not contradict rationality, but can be defined as part of it. Gewirth goes further than this. The aim of his strongly deontological theory is to establish morality, not only as reconcilable with rationality, but as a necessary part of it. Moral soundness and logical flawlessness are equally parts of rationality in the sense that failure in recognizing a justifiable moral demand amounts to a logical mistake in reasoning. And morality is analogous to the criteria of rationality in another sense, too. Gewirth seeks the justificatory basis for morality in a formal device: the aim of moral philosophy is to find a logically necessary, universal principle, which every rational human being must accept, on pain of being logically inconsistent.

Moral philosophy starts with the question: "Why should I be moral?" When this *authoritative* question has been answered in a satisfactory way, the *distributive* and *substantive* questions can be solved as well, that is to say, it becomes possible to explicate the object of moral concern and the content of normative morality.[2] In Gewirth's view, moral philosophy has accomplished its task when the question "Why should I be moral?" has been answered in such a way as to establish a logically necessary supreme moral principle.[3] If there is no such principle morality does not have a unequivocal foundation. The search for right and wrong becomes pointless, and radical relativism remains the only acceptable position.[4] Gewirth claims that his theory shows that morality has, in fact, a rational justification, which all rational agents must accept as valid and obligating, irrespective of their contingent circumstances. In this chapter, I explicate

[1] *Reason and Morality*. The University of Chicago Press, Chicago & London. 1978.

Gewirth's theory concisely, paying special attention to the formal constituents of his theory. Based on this, I analyze the concept of a moral person that is embedded in the theory.

Gewirth builds his argument on a conceptual analysis of action. He holds it possible to show that the consequences of his reasoning are logically necessary, and in this sense binding. In addition to this rational justification of morality, the analysis generates an ultimate moral principle, the universal validity of which is based upon the formal necessity of the argument.[5] Moreover, the theory is designed to bridge both the gap between factual and evaluative statements, and the logical rift between ought-propositions and is-statements.[6]

Why would one want to show that morality is rational, or that it is based upon reason? Does not reason itself need a similar justification? Reason, which comprises both deductive and inductive logic,[7] is the ultimate judge of all validity, Gewirth maintains, because any attack against reason can justify its claims only by relying on reason. Reason is nothing

---

[2] "In the light of these conflicts of criteria and interests, we may distinguish three central questions of moral philosophy. First, there is the authoritative question: Why should one be moral, in the sense of accepting as supremely authoritative or obligatory for one's actions the requirement of furthering or favorably considering the important interests of other persons, especially when these conflict with one's own interests? [...] Second, there is the distributive question: Whose interests other than his own should the agent favorably consider in action? [...] Third, there is the substantive question: Of which interests should favorable account be taken? Which interests are good ones or constitute the most important goods? [...] The various answers to the distributive and substantive questions constitute conflicting criteria of moral rightness, while the authoritative question concerns the reasons for accepting one or another of these criteria as categorically mandatory for actions and institutions." GEWIRTH 1978, 3.

[3] "Now such an answer [to the authoritative question] is obtainable if a supreme moral principle can be shown to be logically necessary so that its denial is self-contradictory." GEWIRTH 1978, 23. "A definitive justification of a supreme moral principle must provide, in a way that does not beg the question, determinate and conclusive answers to the three questions of moral philosophy. The answers must be determinate in that the criteria of moral rightness they establish have certain definite contents such that the opposite contents cannot be derived from the principle that has been justified. The answers must also be conlusive in that they show that these criteria cannot rationally be challenged by any of their competitors and that it is categorically obligatory for all persons to act in accordance with them." GEWIRTH 1978, 21.

[4] GEWIRTH 1978, 7.

that we can intelligibly try to justify, for we cannot even demand a justification without depending on reason. Deductive and inductive inference are simply the only applicable methods for avoiding arbitrariness and subjectivity. Consequently, if we want to have a non-contingent moral theory, we must ground it upon the precepts of reason.[8] Gewirth represents a *criteriological* concept of reason: the formal conditions of correctness in thinking and inferring set the criteria for rational acceptability. Gewirth establishes his moral theory upon this concept of reason, which has the consequence that his moral theory becomes criteriological , too; it determines the rightness and wrongness of action on purely formal grounds.

---

[5] "In this book, [...], I present a new version of a rational justification. The chief novelty is the logical derivation of a substantial normative moral principle from the nature of human action. [...] My main thesis is that every agent, by the fact of engaging in action, is logically committed to the acceptance of certain evaluative and deontic judgments and ultimately of a supreme moral principle, the Principle of Generic Consistency, which requires that he respect his recipients' necessary conditions of action. To prove the thesis, I have argued that the very possibility of rational interpersonal action depends upon the adherence to the morality that is grounded in this principle. Because every agent must accept the principle on pain of self-contradiction, it has a stringent rational justification that is at the same time practical because its required locus is the context of action." Gewirth 1978, x.

[6] Gewirth 1978, 25–26.

[7] "I use 'reason' in a strict sense as comprising only the canons of deductive and inductive logic, including among the latter its bases in particular sense perceptions. [...] Although difficulties may be raised about the general justification of both deduction and induction, in the present context it must suffice to note that, because they respectively achieve logical necessity and reflect what is empirically ineluctable, deduction and induction are the only sure ways of avoiding arbitrariness and attaining objectivity and hence a correctness or truth that reflects not personal whims or prejudices but the requirements of the subject matter." Gewirth 1978, 22.

[8] "It is indeed the case that there have also been historical demands that reason itself in turn pass various justificatory tests set by religious faith, aesthetic rapture, and so forth. But the very scrutiny to determine whether these tests are passed must itself make use of reason. For example, salient powers of reason must be used in order to check whether the products of logical and empirical rationality are consistent with propositions upheld on the basis of faith, or whether the use of reason is compatible with the experiencing of aesthetic feelings, and the like. Thus any attack on reason or any claim to supersede it by some other human power or criterion must rely on reason to justify its claims." Gewirth 1978, 23.

2.3.1. THEORY OF ACTION

The first step in finding a justification for morality is to find its subject matter. Is there anything common among objects of moral concern? Gewirth answers the question affirmatively and maintains that morality necessarily deals with *action*, and more specifically, with the *generic features of action*. These features characterize the genus of action, or its necessary constituents.[9] Gewirth analyzes the concept of action to show that human action has a *normative structure*. By this he means that all action includes implicit evaluative and deontic judgements. These judgements, together with the logical principle of universalizability, then imply a certain normative moral rule.[10] Gewirth exemplifies the role of action in moral theory by likening it to the role empirical data have in natural sciences: action is a basis against which moral rules and empirical statements can be checked for their truth or correctness. Moral judgements can only be true insofar as they correspond with the normative structure of action.[11]

There are two features which are constitutive and therefore logically

[9] "This necessary content [or subject matter] of morality is to be found in action and its generic features. For all moral precepts, regardless of their further contents, deal directly or indirectly with how persons ought to act. The specific modes of action required by different moral precepts are, of course, highly variable. But amid these variations, the precepts require actions; and there are certain invariant features that pertain generically to all actions. I shall call these the *generic features* of action because they characterize the genus or category of action as a whole, as delimited by moral and other practical precepts. Thus, just as action provides the necessary content of all morality, so the generic features provide the necessary content of all action." GEWIRTH 1978, 25.

[10] "[...] because of its generic features, action has what I shall call a 'normative structure', in that evaluative and deontic judgments on the part of agents are logically implicit in all action; a certain normative moral principle logically follows from them. To put it otherwise: Any agent, simply by virtue of being an agent, must admit, on pain of self-contradiction, that he ought to act in certain determinate ways." GEWIRTH 1978, 25–26.

[11] "Since action comprises the factual subject matter of moral and other practical precepts, it serves for moral philosophy a function analogous to that which empirical observational data may be held to serve for natural science: that of providing an objective basis or subject matter against which, respectively, moral judgments or rules and empirical statements or laws can be checked for their truth or correctness. It must be emphasized that this function is only analogous: a moral judgment does not become true simply by stating that some action or kind of action is actually performed." GEWIRTH 1978, 26.

necessary to action. These are *voluntariness* or freedom, and *purposiveness* or intentionality. According to Gewirth's definition, an action is voluntary when it is performed under the agent's uncoerced, informed control, i.e., when it is the object of her free choice.[12] Moreover, an action is intentional if it is performed for some end which constitutes the agent's reason for action.[13] Voluntariness refers to the means of action and purposiveness to its end.[14] Gewirth distinguishes two modes of both voluntariness and purposiveness: they may be either *occurrent*, occurring only at a certain occasion, or *dispositional*, being latent and getting exposed just when the agent considers her ends distinctively and makes a choice concerning an action.[15] Voluntariness and purposiveness embody the mini-

---

[12] Gewirth distinguishes negative and positive conditions for an action to be voluntary. The negative conditions consist in the agent being free of direct external compulsion, whether physical or psychological of nature, the agent acting without being exposed to internal factors affecting her behaviour beyond her control, such as reflexes, ignorance, or disease; and finally that the agent is not being affected by indirect coercing, for example, made to perform a certain action under the threat that a third party will otherwise be hurt. The positive conditions of freedom comprise of the person's being in control of her behaviour by her own unforced and informed choice. All in all, an action's being voluntary involves that the agent's "unforced and informed choice is the necessary and sufficient condition of the behavior. For all behaviors that are the objects of moral and other practical precepts, it is assumed that the persons addressed can control their behaviors in this way. When there is such control, the person chooses on the basis of informed reasons he has for acting as he does." Gewirth 1978, 31.

[13] "By an action's being purposive or intentional I mean that the agent acts for some end or purpose that constitutes his reason for acting; this purpose may consist in the action itself or in something to be achieved by the action." Gewirth 1978, 27.

[14] "Voluntariness refers to the means, purposiveness to the end; voluntariness comprises the agent's causation of his action whereas purposiveness comprises the object or goal of the action in the sense of the good he wants to achieve or have through this causation. Thus voluntariness is a matter of initiation or control while purposiveness is at least in part a matter of consummation." Gewirth 1978, 41.

[15] "We may [...] distinguish [...] different sorts of control and unforced choice on the part of a person that, when applied to his behavior, make it voluntary. (*a*) The control and choice may be occurrent, pertaining directly to the action as it occurs. (*b*) They may be dispositional, in that if the agent had chosen not to perform the action he would not have performed it." Gewirth 1978, 32. "Just as control and choice may be dispositional as well as occurrent, so it is with purposiveness. Agents may not always have purposes clearly in view [...] The purposes for which persons act may be habitual, results of long-standing goal-directed behavior where the goals have ceased to occupy the center of attention." Gewirth 1978, 38.

mal *necessary* and *sufficient conditions* of all rational action. If morality is rationally justifiable, and hence rational, voluntariness and purposiveness must be the minimal necessary and sufficient conditions of moral action as well. Consequently, freedom and intentionality are the universal constituents of human action, its invariable constants, that are independent of the contingent factors of action such as its subject, time or place.[16] This means that nothing qualifies as human action unless it can be characterized as free and intended. Similarly, there is no human agency without voluntariness and purposiveness.

Intentionality means that agents perform their actions for attaining some desired end. To say that an action is intentional implies a positive attitude in the agent towards the goal of her specific action. In this sense, there are no indifferent actions.[17] Intentionality thus forms a connection between an agent's action and her wants or desires. This link is important as a part of Gewirth's moral theory, for he establishes the connection between facts and values on it. A description of an intentional act is a factual statement. This description of facts contains, however, an implicit value judgement in the above sense, and striving for a desired goal presupposes that the agent necessarily values her purposes as good. Hence, a normative proposition is inferred from a factual statement.[18]

Analyzing Gewirth's theory of action closer, we note that his model

[16] "Nevertheless, insofar as they fulfill certain formal conditions, the precepts prospectively assume that the behaviors of the persons addressed will be voluntary and purposive. These conditions, which are found in all moral precepts, consist in certain minimal aspects of rationality and associated normativeness." GEWIRTH 1878, 28. "The moral population addressed by moral precepts varies in part according to the substantive and especially the distributive criteria upheld in different moralities. [...] But however much the persons addressed by various moral and other practical precepts may differ in other respects, it is true of all of them that they are assumed to be able to control their relevant behaviors by their unforced choice for reasons and purposes they can make their own." GEWIRTH 1978, 30.

[17] "What is common to all cases of wanting to do something is that the agent has some sort of pro-attitude toward the purpose of his action. [...] It follows from this that there are no indifferent actions, 'indifferent' meaning that the agent does not care at all whether he performs the action or not. For even if he regards his action as morally indifferent or as not making any difference on some other specific criterion, by the very fact that he aims to do the action he has a pro-attitude toward doing it and hence a positive or favorable interest in doing it." GEWIRTH 1978, 40.

represents the practical syllogism in its classical form.[19] It is the form of intentional human action. How then does moral action relate to other intentional action in Gewirth's theory? Unlike the theorists we have examined so far, Gewirth does not purport to set up further criteria to distinguish morality from the realm of intentionality, but aims to show that all intentional action implicitly includes a normative component upon which morality is based. What does this mean? Does Gewirth's model presuppose that intentionality can be reduced to morality? In order to be able to answer these questions we have to examine Gewirth's theory further.

Gewirth uses his conceptual analysis of action to establish a procedure which he calls the *dialectically necessary method*, for developing moral theory. By calling the method "dialectical", Gewirth conjoins the Socratic-Aristotelian tradition, which sets off from tentative assumptions of the subject matter and proceeds by examining what they imply logically.[20] Gewirth distinguishes his method from another, dialectically contingent method, by saying that the contingent version of the method deals only with variable beliefs, interests or ideals, whereas his own, dialectically necessary method proceeds from statements "necessarily attributable to every agent because they derive from the generic features which constitute the necessary structure of action".[21] The dialectically necessary method only allows propositions, which emerge from the con-

---

[18] GEWIRTH 1978, 41. E.M. ADAMS points out (in his "The Subjective Normative Structure of Agency". *Gewirth's Ethical Rationalism. Critical Essays with a Reply by Alan Gewirth.* Ed. by Edward Regis Jr., The University of Chicago Press, Chicago & London, 1984, 13–16) that by such inference Gewirth actually reduces the meaning of 'good' to 'that which is chosen by the agent', for he does not offer a definition of the good independent of the agent's choices. Gewirth repudiates this criticism by maintaining that in this connection 'good' refers only to what the agent deems to be good, not what really is good. Gewirth does not, however, offer any other means for evaluating the goodness of things independent of their being objects of people's choice: "To hold that these objects are good is to value or prize them, and to say that they are good is to give expression to these attitudes." GEWIRTH 1978, 51–52.

[19] See page 16.

[20] GEWIRTH 1978, 43.

[21] GEWIRTH 1978, 43–44.

ceptual analysis of action or from an agent's necessary beliefs.[22] This method, Gewirth maintains, secures an objective and universally attributable outcome for it is based on the generic features of action and of agency, which are independent of what any agent might regard them to be.[23]

2.3.2. MORAL PERSONHOOD

An agent who manifests her agency by performing the kind of actions defined in the above is:

> a person who is rational in that he is aware of and can give expression to the generic features that conceptual analysis shows to pertain necessarily to his actions, including the logical implications of these features.[24]

That the agent is *rational* means simply that she is consistent or avoids self-contradiction in accepting what her agency involves logically. As a rational agent she is also capable of certain minimal inductive reasoning. The attribute "rational" characterizes any person who controls her

---

[22] "The use of the dialectically necessary method requires a certain sequence of argument. In this sequence, only those propositions are accepted, either as definitively justified or as warranting favorable consideration, that emerge successively from the conceptual analysis of action and of the agent's necessary beliefs. Thus the whole structure of argument leading to the rational justification of the supreme principle of morality will consist only in such rationally necessary propositions." GEWIRTH 1978, 46–47.

[23] "The method I shall use here will be a dialectically necessary one, since this reflects the objectivity and universality reason achieves through the conceptual analysis of action. [...] It is important to note, however, that it is not the dialectically necessary method that determines the generic features of action and hence the general standpoint of agency itself, since the contents of these features are independent of what any agent may think they are. But once these features have been ascertained, as indicated above, the method operates to trace what judgments and claims every agent logically must make from within this standpoint." GEWIRTH 1978, 44.

[24] GEWIRTH 1978, 44.

behaviour for achieving her goals.[25] As we can see, Gewirth defines a moral person in terms of the generic features of action and of agency. So what does this involve? It seems that Gewirth simply *describes* what being a person comprises. But is this all? The real importance of this definition lies in the fact that Gewirth thereby demarcates the morally relevant features from what he regards as ethically irrelevant. Thus, his description of the morally relevant and of a moral person coincide. In relation to this, we must consider whether Gewirth's description of personhood is ethically neutral, or whether it bears normative implications for the theory. Although my critique does not yet permit a full answer to the question, we can already infer something on the basis of what has been said so far.

Gewirth's aim is to establish a moral theory every rational agent must adopt in virtue of her rationality. Let us, for the sake of the argument, accept Gewirth's theory of action and assent to intentionality and freedom as prerequisites of human action and agency. This does not, however, imply that we would have to accept Gewirth's moral theory, because we can always say that we do not agree with his conceptual analysis as a model of the *morally relevant*. We can, on the contrary, legitimately maintain that there are other feasible ways of determining moral significance. We can even claim that Gewirth's conceptual paradigm is defective, because descriptions which fall short of the given conditions receive no moral consideration.[26]

As we have already noticed, the purposiveness of action contains an implicit evaluative proposition. This proposition states that the purpose, which is both the object of the agent's desire and the reason for her acting, appears to her as good. The goodness of the desired object rests on some (implicit) criterion the agent holds in regard to her aims.[27] Now,

---

[25] "I shall henceforth refer to the agent who grasps or accepts such entailments [logical relations between propositions and simple and direct entailments from beliefs of respective propositions] as a *rational agent*. It is to be noted that the criterion of 'rational' here is a minimal deductive one, involving consistency or the avoidance of self-contradiction in ascertaining or accepting what is logically involved in one's acting for purposes and in the associated concepts. In addition to such deductive rationality, a certain minimal inductive rationality may also be attributed to the rational agent." Gewirth 1978, 46.

the implicitly evaluative nature of action implies a further proposition. As the agent necessarily attaches a positive valence to her particular goal, so she similarly values the features which constitute her action, i.e., her freedom and intentionality. In this way, voluntariness and purposiveness are generic and, hence, necessary goods of every agent. An agent must value her freedom and intentionality, no matter what her specific goal is, because they are necessary prerequisites for her to attain whatever she wants to acquire. This fact reveals the basic nature of these goods.[28] Furthermore, the generic goods are also central to an agent's well-being because they constitute her capability of action.[29]

The argument concerning freedom and intentionality as necessary goods is central to Gewirth's theory. He maintains as follows:

> Since the agent regards as necessary goods the freedom and well-being that constitute the generic features of his successful action, he logically must also hold that he has rights to these generic features, and he implicitly makes a corresponding right-claim.[30]

---

[26] This interpretation gets support from GEWIRTH's text (1978, 29): "The analysis of the concept of action is not to be regarded as yielding results that are merely 'conceptual' as opposed to 'real'. Rather, the concern is to differentiate, from the many and varied real features of human behaviors, those that *constitute human action in the relevant sense*. What the utterers of such precepts prospectively envisage about future actions falls within the real possiblities of human conduct." (Emphasis added.) See also GEWIRTH 1978, 47: "If at any stage of this sequence [within the dialectically necessary method], some particular agents or groups of agents have contingent beliefs or principles that are opposed to such propositions, this opposition will carry no justificatory weight. For a main point of confining the argument to rational necessities is to attain an objective standpoint from which such beliefs or principles can be critically evaluated. The dialectically necessary method attains such a standpoint *by restricting the propositions it admits as justified* to those that follow from the concept of action, including the beliefs or judgments every agent is necessarily warranted in having on the basis of his acting for purposes he wants to achieve." (Emphasis added.)

[27] "In acting, the agent envisages more or less clearly some preferred outcome, some objective or goal he wants to achieve, where such wanting may be either intentional or inclinational. He regards this goal as worth aiming at or pursuing; for if he did not so regard it he would not unforcedly choose to move from quiescence or nonaction to action with a view to achieving the goal. This conception of worth constitutes a valuing on the part of the agent, he regards the object of his action as having at least sufficient value to merit his acting to attain it, according to whatever criteria are involved in his action." GEWIRTH 1978, 49.

From the fact that an agent regards the prerequisites of action, which constitute her moral personhood, as her necessary goods, it follows that the agent reasons that she has a *right* to them. Namely, if a rational agent is to claim any rights at all, there cannot be a more urgent object for her right-claim than the goods necessary for her agency. Additionally, there cannot be a stronger justificatory basis for anyone's claim on something as her right than that the object of her claim is the necessary condition of her agency.[31] Moreover, an agent's description of herself as an intentional agent is both a necessary and a sufficient condition for the justification with which she supports her right-claim as an agent. It is a necessary condition for the reason that every agent performs her actions for attaining purposes she esteems as good; and it is a sufficient condition on the

---

[28] "The agent's positive evaluation extends not only to his particular purpose but also a fortiori to the generic features that characterize all his actions. These features constitute, in his view, what I shall call *generic goods*. Since his action is a means of attaining something he regards as good, even if this is only the performance of the action itself, he regards as a necessary good the voluntariness or freedom that is an essential feature of his action, for without this he would not be able to act for any purpose or good at all." Gewirth 1978, 52. "In addition to the voluntariness or freedom of his actions, the agent also values their generic purposiveness as a necessary good." Gewirth 1978, 53. Gewirth divides the good included in purposiveness into three kinds of goods which he calls basic goods, non-subtractive goods, and additive goods respectively. The basic goods include the physical and psychological dispositions necessary for action; the subtractive goods comprise the agent's retaining and not losing whatever she already has and which she regards as good; and the additive goods consist in the objects of the agent's purposive action. Gewirth 1978, 53–56.

[29] "What I have tried to show, then, is that all purposive action is valuational, and that agents regard as good not only their particular purposes but also the voluntariness or freedom and purposiveness that generically characterize all their actions." Gewirth 1978, 57. "In any case it is not the particular purposes and outcomes but rather the generic abilities and conditions that for the agent primarily constitute his well-being since they are the necessary conditions of all his pursuits of his purposes." Gewirth 1978, 61.

[30] Gewirth 1978, 63.

[31] "If a rational agent is to claim any right at all, could anything be a more urgent object of his claim than the necessary conditions of his engaging both in action in general and in successful action?" Gewirth 1978, 63. "[...] what greater justification could any person have for claiming any rights relevant to action than that their objects are necessary for his engaging in any purposive actions at all or for his succeeding in any such actions?" Gewirth 1978, 72. "[...] he is entitled to freedom and well-being because of the genuine necessity, generality, and fundamental character of the justifying reasons on which his claim is based." Gewirth 1978, 73.

basis that intentional actions become impossible without the constitutive prerequisites of action.[32]

The form of this implicit right-claim is negative, as it presupposes that other persons ought at least to refrain from interfering with the conditions necessary for a person's agency.[33] This involves, according to Gewirth, that action has a *deontic structure*: the agent, in her action, implicitly claims to have a right to the generic features of her action, from which it follows that others must respect this right by not depriving her of the constituents of her agency. Accordingly, Gewirth calls the conditions that secure the agent's freedom and purposiveness *generic rights*.[34] These rights are the prerequisites and the grounds for all other rights, and in this capacity they are fundamental rights proscribed to all human beings on the basis of their agency.[35] The deontic structure of action

[32] "The agent's description of himself as a prospective purposive agent is both a necessary and a sufficient condition of the justifying reason he must adduce for his claim to have the generic rights. That it is a necessary condition can be seen from the fact that every agent performs his actions by virtue of having purposes whose fulfillment he regards as good. And it is because freedom and well-being are required for such purposive actions that every agent claims the rights to these generic features of action. [...] This description of the agent is also a sufficient condition of the justifying reason he must adduce for his having the generic rights." GEWIRTH 1978, 109–110.

[33] "If he regards these conditions as indeed necessary for the very possibility of his agency and for his chances of succeeding in his actions, then must he not hold that all other persons ought at least to refrain from interfering with the conditions? Since this 'ought' entails correlative rights insofar as it signifies what the agent regards as his due, the latter question may also be put in the following equivalent form: Must not the agent hold that he has rights to these necessary conditions of his agency?" GEWIRTH 1978, 64.

[34] "This second affirmative answer — that every agent must hold that he has rights to the necessary conditions of agency, freedom and well-being — entails that action has a deontic as well as an evaluative structure. Through its deontic structure, action encompasses not only the agent's evaluative judgments about the necessary goodness of his having freedom and well-being but also deontic judgments he makes or accepts that he has rights to these generic features of action. I shall hence call them *generic rights*." GEWIRTH 1978, 64.

[35] "These rights are also generic in one or another of two further senses: either they subsume other rights in that the others are specifications of the rights to freedom and well-being, or they take precedence over other rights in that the latter, if they are to be valid, must not violate the rights to freedom and well-being. In these respects, they may be called 'fundamental rights'. They are also constitutive rights in that their objects are the proximate necessary conditions of all agency. And they are 'human rights' in that they are rights that all humans have as human agents [...]" GEWIRTH 1978, 64.

implies that an agent has a right to the generic features of her action against other persons for the reason that these features are the necessary conditions of her agency.[36] In addition to these negative rights, which agents have because they are agents, the deontic structure of action implicitly entails a positive right as well. According to Gewirth, other persons have a positive duty to assist the agent if she lacks the necessary constituents of agency.[37]

Before moving further, we must ask what actually takes place in Gewirth's analysis of the implicitly deontic structure of action. Perhaps the most significant feature is the shift from conceptual prerequisites of free and purposive action to *human rights*. Here we encounter a move from a description the agent uses to designate her action to a moral claim she directs towards others on the basis of her moral personhood. What does this shift actually involve? Is this a legitimate move? Before answering these questions, let us first examine what significance Gewirth himself gives to his investigation.

The analysis of action shows, Gewirth maintains, that action has an implicit evaluative as well as a deontic structure. This conceptual examination helps us to build a link between the factual and the normative propositions, and it clarifies the conditions on which we may say something normative of factual reality. Both the gaps between fact and value, and "is" and "ought" are thus bridged.[38]

To raise a question of the legitimacy of Gewirth's move, we can accuse Gewirth's conceptual analysis of being question-begging: he

---

[36] "The agent holds that other persons owe him at least noninterference with his freedom and well-being, not because of any specific transaction or agreement they have made with him, but on the basis of his own prudential criteria, because such noninterference is necessary to his being a purposive agent." GEWIRTH 1978, 66. "The final ground for maintaining that the agent must hold that he has rights to the generic goods of freedom and well-being is that, unlike the particular goods on purposes for which he may act, the generic goods are the necessary conditions not merely of one particular action as against another but of all successful action in general." GEWIRTH 1978, 77.

[37] "[...] the agent's right-claim also entails, in a secondary way, that under certain conditions other persons ought to assist him to have freedom and well-being. These conditions occur when failure to give such assistance would result in his failing to have these necessary goods." GEWIRTH 1978, 67.

attaches moral concepts to the agent, and thus makes his argument circular. He tries to give a rational justification to morality by using an argumentation which is already moral. Against such objections Gewirth maintains that the "ought" and "rights" used here are not moral but purely prudential concepts, because the deontic concepts do not, in this context, refer to other persons' *interests* but solely to the constitutents of action.[39] The move from the prudential to the moral "ought" takes place first when an agent comes to notice that she cannot avoid admitting that every other agent is entitled to the same basic rights as she herself on the very ground she claims these rights to herself, namely, on the ground that they, too, are agents.[40] The reason why no agent can deny the basic rights from other people is that the same description, "this is an agent", applies to all people, and not only to the agent herself.[41] To deny this would mean breaking the logical rule of universalizability, and hence, contradicting oneself.[42] Gewirth further maintains that although the principle of universalizability has such a central position in the move from the non-moral to the moral realm, the principle falls short of being a substantive moral rule in this context.[43] It becomes a moral principle only when the universalization concerns a moral property, i.e., when the

---

[38] "[...] the analysis of action shows how, beginning from a descriptive concept and a factual statement, evaluative and deontic judgments can be logically derived therefrom. [...] An important consequence of these considerations is that in the logical structure of action both the gap between fact and value and the gap between 'is' and 'ought are bridged." GEWIRTH 1978, 102.

[39] "The concepts of 'rights' and 'ought' as here invoked by the agent are not, however, moral. [...] Now the criteria or grounds to which the agent appeals to justify his having the generic rights are, so far, not moral ones: they do not refer to the most important interests of at least some persons other than the agent." GEWIRTH 1978, 69. GEWIRTH (1978, 1) defines morality as " a set of categorically obligatory requirements for action that are addressed at least in part to every actual or prospective agent, and that are concerned with furthering the interests, especially the most important interests, of persons or recipients other than or in addition to the agent or the speaker." This definition implies that, according to Gewirth, only interpersonal issues belong to the moral realm. This also means that one cannot harm oneself morally.

[40] "[...] it can be shown by a further dialectically necessary argument that the agent must admit that such rights equally belong to all other prospective human agents. By admitting this, he is logically committed to accept the supreme principle of morality." GEWIRTH 1978, 103.

property which is being attributed to more than one object, on the basis of their being similar in regard to a relevant feature, is deontic or moral.[44] That the principle is only logical and not morally normative is further explained by the fact that the principle sets no criteria for the relevant similarities on the basis of which the universalization is to take place.[45]

Is Gewirth's argument satisfactory? Does he escape the criticism that his argument is circular, deriving moral concepts from terms that are only seemingly non-moral? He does not: Gewirth's theory does not bridge the gap between "is" and "ought" any more than it establishes a

---

[41] "Now whatever the description under which or the sufficient reason for which it is claimed that a person has some right, the claimant must admit, on pain of contradiction, that this right also belongs to any other person to whom that description or sufficient reason applies." GEWIRTH 1978, 104–105. "Since the agent must hold that he has the generic rights for the sufficient reason that he is a prospective purposive agent, he must admit that all prospective purposive agents have these rights." GEWIRTH 1978, 127. According to Gewirth, the crucial factor in other persons' being agents is not whether they actually can act as agents but their purposiveness: "While it is true that to act requires certain abilities, what is crucial in any agent's reason for acting is not his abilities but his purposes." GEWIRTH 1978, 124. Consequently, the generic rights belong, not just to fully capable human adults, but to children, mentally handicapped and demented people as well in so far as these people are capable of the attainment of agency; see GEWIRTH 1978, 141–142.

[42] "This necessity is an exemplification of the formal principle of universalizability in its moral application, which says that whatever is right for one person must be right for any similar person in similar circumstances. But this formal moral principle, in turn, derives from a more general logical principle of universalizability: if some predicate P belongs to some subject S because S has the property Q (where the 'because' is that of sufficient reason or condition), then P must also belong to all other subjects S_1, S_2, [...], S_n that have Q. If one denies this implication in the case of some subject, such as S_1 that has Q, then one contradicts oneself." GEWIRTH 1978, 105.

[43] "The principle of universalizability even in its moral application is not itself a substantial normative moral principle, not only because, depending on the criterion it uses for relevant similarities or for the property Q, it gives results that are morally quite diverse and even opposed to one another, but also because it simply explicates what is involved in the concept of 'because' as signifying a sufficient reason. Hence, in using the principle of universalizability to establish the supreme principle of morality I shall not be using a substantial moral principle." GEWIRTH 1978, 105.

[44] "The logical principle of universalizability is given a deontic or moral application by interpreting the predicate P in the above pattern as a deontic or moral predicate. It may then be formulated, among other ways, as follows: if one person S has a certain right because he has quality Q [...], then all persons who have Q must have such a right." GEWIRTH 1978, 106.

link between facts and values. Even if we could not find fault with Gewirth's conceptual analysis or his logical inferences, we can point out two facts which make Gewirth's justificatory programme unsustainable. First, the fact that we can logically infer something from a set of premises does not constitute the truth or acceptability of these premises. We can make logically correct inferences from false premises, and accordingly, accept Gewirth's conclusions as logically valid but still abandon them on the ground that we do not acknowledge that the premises are true. Second, the implications which Gewirth derives from the concept of moral agency are not purely logical. As I have pointed out earlier,[46] moral agency and moral personhood define the scope of the morally relevant in Gewirth's theory. Gewirth simply presupposes that we accept the definition of moral personhood as given, for there is no rational justification for that in the theory. Still, seeing the moral person primarily as an acting agent, valuing the conditions of her action as a prerequisite of everything else, may not be ethically as neutral as Gewirth seems to presuppose.

We can now answer the questions raised above concerning the distinction between moral action and other intentional action.[47] Should we accept Gewirth's differentiation between the moral and the non-moral "ought" and "rights"? According to him, the distinction between morality and intentionality in general lies in there being morally relevant similarities between the agent as moral subject and other people as moral objects. Only in such cases can the necessary universalization have moral significance. Gewirth does not, however, explicate what these morally relevant similarities are.

---

[45] "The apparently egalitarian import of the moral principle of universalizability is severely restricted by the fact that the principle allows complete variability with respect to content. One kind of contentual variability is that the actions it is right to perform, according to the principle, may vary indiscriminately in accordance with the variable inclinations or ideals of agents; this violates the requirement of categoricalness for a supreme moral principle." GEWIRTH 1978, 106.

[46] See page 157.

[47] See page 155.

2.3.3. The principle of general consistency

Acting agents become moral subjects as they come to occupy morally relevant roles, or in other words, intentional action becomes morally relevant when agents are involved in actions which *affect the interests of other people*. Gewirth calls such interpersonal actions *transactions* and the persons affected by an agent's action *recipients*. It follows from the conceptual features of action that an agent is logically committed to respect the generic rights of her recipients in virtue of their being prospective purposive agents. This entails that an agent has a negative duty not to interfere with the recipient's right to the conditions constitutive for her agency; a recipient must at least be free to choose whether she wants to participate in the transaction or not.[48]Consequently, every agent has a negative obligation and a positive duty towards her recipients. First, an agent must acknowledge the obligation to refrain from depriving other persons of the basic conditions of their agency. Second, an agent has to yield to the duty to assist other persons to have freedom and well-being when the lack of these prevents them from acting as agents.[49] The general principle which expresses these obligations is the *Principle of General Consistency*, or *PGC*: "Act in accord with the generic rights of your recipients as well as of yourself."[50] The *PGC* thus combines the formal consideration of consistency (the principle of universalizability) with the

[48] "Since the recipients of the agent's action are prospective agents who have purposes they want to fulfill, the agent must acknowledge that the generalization to which we saw that he is logically committed applies to his recipients: they too have rights to freedom and well-being. Their right to freedom means that just as the agent holds that he has a right to control whether or not he will participate in transactions, so his recipients have the right to control whether or not they will participate. Hence, the agent ought to refrain from interfering with their freedom by coercing them: their participation in transaction must be subject to their own consent, to their own unforced choice." Gewirth 1978, 134.

[49] "It follows from these considerations that every agent logically must acknowledge certain generic obligations. Negatively, he ought to refrain from coercing and from harming his recipients; positively, he ought to assist them to have freedom and well-being whenever they cannot otherwise have these necessary goods and he can help them at no comparable cost to himself." Gewirth 1978, 135. See also Gewirth 1978, 137.

[50] Gewirth 1978, 135.

material consideration of rights (agents' rights to basic freedom and well-being) to the generic features of goods of action (freedom and purposiveness).[51] Consequently, as a moral principle the *PGC* is necessary in two ways: it is logically necessary, while violating it means contradicting oneself; and it is materially necessary, or categorical, because no agent can escape its obligation by simply changing her inclinations, interests, or ideals.[52] As a material moral principle the *PGC* is *egalitarian* and *universalist* since it requires an equal distribution of the most general rights of action.[53] Gewirth maintains that the *PGC* is a rationally grounded moral principle, because its rationality is based on the preceding conceptual analysis. This means that there is now a rationally justified substantive moral principle which has been deduced from rational considerations about the features that necessarily pertain to all action.[54] This implies that an agent, if she acts rationally, may not break with this principle.[55]

To classify the morality implied by the *PGC*, we can say that it is *negative* and *minimalist*. It is negative, because the obligation of non-interference is primary to any positive duty. The agent has a moral duty to refrain from inhibiting others to act as free intentional agents in order to secure for herself the possibility to such agency. The principle is minimalist, for the only positive duty it ascribes is that others must be helped to acquire the conditions of agency, so that they can pursue their own goals. The *PGC* does not obligate the advancement of other people's

---

[51] GEWIRTH 1978, 135.

[52] "The *PGC* is a necessary principle in two ways. It is formally or logically necessary in that for any agent to deny or violate it is to contradict himself, since he would then be in the position of holding that rights he claims for himself by virtue of having certain qualities are not possessed by other persons who have those qualities. The principle is also materially necessary, or categorical, in that, unlike other principles, the obligations of the *PGC* cannot be escaped by any agent by shifting his inclinations, interests, or ideals, or by appealing to institutional rules whose contents are determined by convention." GEWIRTH 1978, 135.

[53] GEWIRTH 1978, 140.

[54] Gewirth 1978, 148.

[55] "It is important to note here a certain connection between action and judgment. When an agent violates the *PGC* intentionally infringing a generic right of his recipients, he in effect denies that they have this right and he thereby ceases so far forth to be rational." GEWIRTH 1978, 139.

substantive goods, it presents only a duty to advance the necessary means for realizing such goods.

We have now accomplished the presentation of Gewirth's theory, but before closing the subject I wish to take up two cases for clarifying my suggestion that Gewirth's choice of basic concepts may not be morally neutral, but that these notions presuppose a certain moral position. The examples concern Gewirth's view on agents who are not fully competent of rational action, and his discussion on violation against the criteria of rationality. First, Gewirth maintains that agents who are defective in their ability to make full use of the necessary prerequisites of action are entitled to that amount of generic goods as they are capable of using.[56] Additionally, they must be treated by others so as to "effect whatever improvements may be possible in the direction of *normal agency*" (my emphasis).[57] Gewirth's formulation shows that his analysis of action forms a *norm* in the sense that it describes human action through its essential features, which, again, serve as criteria for what can be considered as human action. What "defective" humans do is relevant, from the perspective of moral theory, to the degree to which their actions square with the generic features of action defined by the conceptual analysis. Gewirth claims that his analysis is conceptual, revealing the logical implications of being an agent, but he does not notice that it becomes normative in the sense that it selects the criteria along which human activities are estimated, or a point of view through which human life is regarded. For this reason, it is legitimate to say that the choice of basic concepts has a normative bearing Gewirth does not pay attention to.

My second remark concerns the fact, acknowledged by Gewirth, too, that agents may disregard the demands of rationality in their action. Gewirth does not discuss this problem further, he simply states that this is a fact to which rational agents must pay attention.[58] If we follow Gewirth's line of thought, the validity of the *PGC* is based upon logical necessity. But, we can ask, whether breaking with the *PGC* is similar to,

[56] Gewirth 1978, 141–142.

[57] Gewirth 1978, 142.

[58] Gewirth 1978, 140.

say, violating the law of contradiction? It is difficult to see how this could be the case for, even if it is difficult, we can imagine what our lives would be like if we constantly violated the *PGC*, whereas it is inconceivable even to think what continuous violation of the basic logical principle of contradiction would involve.[59] The former could be unbearable, but it would not be comparable to violating basic logical principles.[60] While we would call those who refuse to obey the *PGC* immoral, or amoral, those who consequently break with logical consistency (if there even can be such people) are in our eyes incomprehensible. This shows that the *PGC* is not a logically grounded principle in the sense Gewirth claims it to be.

[59] Here Gewirth's position resembles that of Hare's, see page 58.

[60] See, e.g., Nielsen's (1984, 68–69) example of an amoral ganster who does not find violating the Gewirthian principles at all difficult and still acts successfully.

2.4. The contractarian person

The mutual starting point of the three theories studied in this chapter is intentional action, seen from the perspective of an acting agent. It is crucial for any acting agent that the prerequisites of her intentional action are satisfied. Unqualified intentional action becomes moral when what the agent does has an effect on the conditions of other intentional agents' action, that is, when intentional action takes place in a social context. The agent whose action has this effect is the moral subject and those whose conditions of action are affected are moral objects. Consequently, the necessary conditions of intentional action in a social situation define what these theories regard as morally relevant. It is characteristic of this definition that it is strictly formal, concentrating on the conceptual prerequisites of action without any reference to a goal.

The definition of the morally relevant has a normative impact on these theories because it determines the focus of normative morality. The task the ethicists of this chapter give to their moral theories is to offer a procedure for eliminating actions which deprive others of the necessary conditions of intentional action. Their approach means that morality is primarily understood as a *restriction* which is formal in nature being based on the conceptual features of intentional action. Accepting morality as a constraint to action guarantees a person, both as a moral subject and as a moral object, the necessary conditions of intentional action. This also means that confining oneself to morality in a community of similar agents secures the possibility of rational action and plans for life.

The analysis revealed a feature we have noticed earlier in the context of the utilitarian theories we have studied: here, too, the definition of the morally relevant accords with the moral person. The moral person is defined by the universal characteristics of any morally relevant situation and accordingly she lacks all particular qualities whether permanent personal characteristics or occurrent desires. This concept of a moral person represents a model everyone must accept by virtue of being an

intentional agent who strives to gain something through action. Furthermore, the idea in these theories is that this definition of the moral agent makes it possible for every rational agent to identify herself with the formal features of this moral person, whatever her actual particular aims are. The concept of a contractarian moral person thus represents universalizable formal features that are also particularly and individually acceptable.

Unlike the utilitarian theories the theories examined in this chapter make a clear distinction between the moral and the non-moral. Morality is conceived as a restriction, not as a maximizing procedure. This means that these theories represent a minimalist conception of morality: its primary function is to prevent actions which the theory defines as morally incorrect. It is not the task of morality to produce good. The clear distinction between the moral and the non-moral can also be seen in the contractarian concept of a person. Thus, it is possible to make a distinction between the moral person and the non-moral or the natural person in these theories. We have already pointed out that the necessary conditions of intentional action define the moral person. The natural person, for her part, is a rational agent in the sense that the aim of her action is the realization of her desired goals. Her goals, however, are determined by her particular, individual and contingent desires, by her life history, social background, and the like. In other words, the natural person represents any (normal) actual person.

Having the clear distinction between the moral and the non-moral person in mind we can now ask what is the relationship between the formal, universalizable moral person and ourselves as natural persons, as individuals with our particular characteristics. What purpose does this distinction serve in the theories? We have already noticed that this feature is connected with the minimalist nature of contractarian morality and with the conception of morality as a restriction. There is, however, a third element in the theories which is relevant in this connection. The distinction between the moral and the natural person becomes understandable given that the contractarian moral theories attempt to give a rational justification for the institution of morality. This justification can

only make sense if it is presented to the natural, non-moral agent, for to present it to the moral agent would mean accepting the justification a priori. The non-moral person must accept moral constraint because she is an intentional agent and any constraint is presented as a necessary condition for acting as such an agent. Accepting constraint essentially involves adopting the role of the moral person in any morally relevant situation. The natural person realizes her intentional agency by acting as if the characteristics of the moral person were her own.

3. *Virtue theories*

In his book *Contemporary Moral Philosophy*[1] G. J. Warnock introduced the most important lines of moral thought in the twentieth century, but what is remarkable in this context is that he does not make a single reference to any theories in which virtues would play even a minor role.[2] Since the late 1960's the situation in the field of moral theory has radically changed. An introduction to modern ethical thinking, if written at present, could not possibly leave out the branch of ethical thought that is widely known as virtue ethics.[3] Ethics of virtue[4] has by now become a major alternative to the utilitarian and deontological theories of the first part of this century.[5] This does not necessarily mean that we could regard virtue ethics as a fully developed, alternative *ethical theory*; we

---

[1] G.J. Warnock, *Contemporary Moral Philosophy.* MacMillan, London, 1967, reprinted in 1981.

[2] Warnock groups contemporary moral theories (in the English-speaking world) with intuitionism, emotivism, and prescriptivism. In his presentation Warnock is very critical of all these varieties of moral theory: "Much recent moral theory has been misguided in its aims and unrewarding in its results". Warnock 1967, 76. The remedy for the inadequacy of these theories is, however, not to be sought from somewhere outside the philosophy of language; modern moral theories have not gone astray because of their study of the meaning of moral terms, on the contrary, this is just what we have to do in order to clarify what we deal with when we speak about moral matters. Warnock 1967, 75–76. Nothing is said about virtues in another work of the same period, either, namely in *Theories of Ethics*, ed. by Philippa Foot. Oxford. 1967. The volume was published in order "to bring together important recent writing in major areas of philosophical inquiry" (according to the back cover of the book), which, in this case, is ethics. It is noteworthy how the editor Philippa Foot characterizes the book: "The articles reprinted here centre round two topics lately much debated: firstly the nature of moral judgement, and secondly the part played by social utility in determining right and wrong. Both these debates go back to the eighteenth century, for at that time philosophers divided for and against the moral sense and intellectualist theories of moral judgement, and at the end of the century Bentham laid down that the principle of utility was the foundation of moral good." *Theories of Ethics*, 1967, 1. Questions concerning the virtues receive no attention.

might rather have to understand it as a newly revived *approach* to ethical thought, neglected for a long time in moral philosophy.[6]

Although the interest in virtues in modern moral philosophy is recent, we can, nevertheless, distinguish two phases. From the late sixties to 1981, the focus of the approach centred round criticism directed against mainstream theories, and virtues were offered as a tentative alternative to the more established ways of moral theorizing. Including virtues in moral theory would help people avoid certain of the mistakes of the dominant theories. The analytical approach was criticized for its narrowness: it focused attention on the conceptually analyzable properties of separate actions, instead of examining the moral agent. Despite their criticism, these writers had not yet developed any methodological alternatives for conducting moral philosophy.

---

[3] A good example of the new importance the study of virtues have gained within a relatively short period are the editions of William K. Frankena's book *Ethics* (1963 and 1973 respectively). In contrast to the first edition, which only brief mentions the virtues (see FRANKENA 1963, 49–51), the revised version gives substantial attention to the ethics of virtue and related themes, (see FRANKENA 1973, xvi, 63–67, and also the section 'The good life', pp. 92–94). To justify taking up this point of view Frankena writes: "The notion of an ethics of virtue is worth looking at here, not only because it has a long history but also because some spokesmen of "the new morality" seem to espouse it." FRANKENA 1973, 63. See also David MCNAUGHTON's *Moral Vision: An Introduction to Ethics*. Basil Blackwell. Oxford and New York. 1988. Although this book mainly deals with the realism – non-cognitivism debate, a section taking account of moral virtue is included. It seems that questions concerning the virtuousness of the moral agent are of such importance that they cannot be excluded without the exposition of ethics remaining incomplete. MCNAUGHTON 1988, 115–117. See also Gregory E. PENCE's article "Recent work on virtues". *American Philosophical Quarterly*. 1984. 281–297.

[4] The theories under scrutiny in this chapter fit the general characterization of either "virtue ethics" (see, e.g., LOUDEN 1990, TRIANOSKY 1990), "virtues ethics" (see, e.g., CLOWNEY 1990), "ethics of virtue" (see, e.g. FRANKENA 1973), or even neo-Aristotelianism (see, e.g., GEACH 1977). In this study I use the terms "virtue ethics" and "ethics of virtue" interchangeably to designate these theories.

[5] See, e.g., TRIANOSKY 1990, 335; CLOWNEY 1990, 50. LOUDEN (1990, 93–94) finds it even appropriate to state: "It now seems safe to say that the genre has not only established itself as a settled paradigm within ethics, but that it is on the verge of achieving hegemony as the outlook of choice among younger writers in ethics. Today one finds fewer and fewer theorists engaging in efforts to construct viable utilitarian or deontological systems: Aristotle has replaced Mill and Kant as the classical moral philosopher most likely to inspire allegiance."

[6] See especially LOUDEN 1984, 227; and LOUDEN 1990, 94.

The first advocates for the importance of virtues were Elisabeth Anscombe, G. H. von Wright and Iris Murdoch.[7] Peter Geach represents a version of Catholic moral philosophy in which the classical cardinal virtues along with the theological virtues are given a central role.[8] Philippa Foot refers to virtues as the only sound starting point for moral philosophy,[9] but she does not aim at developing a full moral theory on virtues. James D. Wallace is the first to present a more elaborated and detailed approach to the theme, but he does not yet present his version of virtue ethics as a theoretical alternative.[10]

The turning point for virtue ethics has been, however, Alasdair MacIntyre's book *After Virtue* published in 1981.[11] In his book MacIntyre not only promotes virtue ethics as a substantive moral theory, but, more importantly, he suggests that the prevailing moral philosophy has to alter the theoretical apparatus with which it examines moral questions. Since that time, other writers have joined the discussion with their own analogical demands.

As a substantive moral theory virtue ethics priorizes questions regarding the good instead of those of the right. According to its view, the basic judgements of morality are not deontic but aretaic in kind and they concern virtues in lieu of principles or rules. Further, the moral quality of an agent, rather than that of a deed, receives prior attention in their ethical evaluation.[12] These theories typically place the question of the good in a

---

[7] G.E.M. ANSCOMBE's famous article "Modern Moral Philosophy", *Philosophy*, 33, pp, 1–19, 1958, has been regarded as one initiator for the new interest in virtues; see, e.g., PENCE 1984, 281; LOUDEN 1984, 227–228. Iris MURDOCH, *The Sovereignty of Good over Other Concepts*. The Leslie Stephen Lecture 1967. Cambridge University Press, Cambridge, 1967; G.H. VON WRIGHT, *The Varieties of Goodness*. Routledge & Kegan Paul, London, 1968.

[8] Peter GEACH, *The Virtues*, Cambridge, London, New York, Melbourne, Cambridge University Press, 1977.

[9] Philippa FOOT, *Virtues and Vices and Other Essays in Moral Philosophy*. Basil Blackwell, Oxford, 1978. See FOOT 1978, xi.

[10] James D. WALLACE, *Virtues and Vices*, Cornell University Press, Ithaca and London, 1978. In the present study I refer to the 1986 impression.

[11] Alasdair MACINTYRE, *After Virtue: A Study in Moral Theory*. Duckworth, London, 1981. In the present study I use the second (1987) edition of the book, which incorporates MacIntyre's comments on criticism directed against his ideas.

wider context, namely, that of a human life. Moral virtues, which are seen prior to ethical principles, are regarded, not solely from an instrumental point of view, but as values constituting the good life. As such these theories are pluralistic: there are many values and they cannot be reduced to a single super-value but have an independent worth. During what I have called the second period, virtue ethicists have also revived the study of classical and medieval moral theories;[13] and in this connection it is interesting to ask what has actually motivated this revival. Has the interest in historical texts only a recording role, or is it connected with the attempt to introduce a new theoretical understanding about what is involved in morality?

The present chapter examines the two phases of virtue ethics respectively. The first part of the analysis concentrates on the theories of Philippa Foot and James D. Wallace. The second part of the present chapter then introduces the theories of Alasdair MacIntyre, Martha Nussbaum and Charles Taylor.

3.1. The rediscovery of virtues

Philippa Foot and James D. Wallace both introduce virtues as an improvement in comparison to prevailing positions in ethics.[14] Virtue-based morality is a welcome replacement for the non-cognitivist, emotivist and Kantian approaches to ethics. There are two particular issues of interest in the present connection, namely the meaning of moral lan-

[12] See, e.g. FRANKENA 1973, 63–65; LOUDEN 1984, 228–229.

[13] A huge amount of literature has appeared during the last two decades on classical virtue-centred ethics, see, e.g., NUSSBAUM 1986; TAYLOR 1989; MACINTYRE 1987, 1988; *Essays on Aristotle's Ethics.* Ed. by Amélie Oksenberg RORTY. University of California Press, Berkeley, Los Angeles, London, 1980; *Ethical Theory: Character and virtue.* Midwest Studies in Philosophy vol XII 1988; HUDSON 1986; KENNY 1978; RORTY 1988; SHERMAN 1991.

guage and the nature of morality as an institution. Both Foot and Wallace maintain that the sharp distinction which has been made between facts and values since Hume is misguided, and that as long as we keep up this division we cannot understand the nature of moral terms.Foot has developed a detailed criticism against the distinction between facts and values, so I will follow her discussion of the topic and only occasionally refer to Wallace.

Hume has initiated a line of thought in ethics which is still influential, as the popularity of emotivism and prescriptivism show. According to Foot, two presuppositions characterize the original, as well as the present forms of this position. First, a strict separation is made between statements concerning facts on the one hand, and those expressing value, on the other. Second, it has been presupposed that moral statements are best understood as expressions of subjective approval or disapproval.[15] These false views are connected with each other and should both be abandoned. Facts and values are not really separable but there is a link between them. In relation to this, Foot also aims to demonstrate that moral judgements are not expressions of our feelings and sentiments but

---

[14] In my presentation I will concentrate on Philippa FOOT, *Virtues and Vices and Other Essays in Moral Philosophy*. Basil Blackwell, Oxford. 1978 and James D. WALLACE, *Virtues and Vices*. Cornell University Press, Ithaca and London, 1978. I will also occasionally refer to Peter GEACH's *The Virtues*. Cambridge University Press, Cambridge, London, New York, Melbourne, 1977; to make necessary comparisons to a recent, but traditionally Catholic model of virtue ethics.

[15] Hume's ethical theory is subjective, because his theory allows no objective criteria for calling something good or bad. A statement attributing a moral quality to an object does not say anything about the object, it simply utters the speaker's sentiment towards the object in question. Hence, a moral proposition only expresses something of the evaluating agent, not of the thing of which it is stated. Hume was led to his idea of total separation between facts and values by his thought that only desires and passions can excite action, whereas reason is always inert or passive in this respect. The connection between reason and the chosen action is purely contingent, depending on the direction of the real initiator of the action, the desire. Hume's solution opens a gap between facts and values, and denies all logical connections, or connections of meaning between moral approval and the objects of moral approval. The criteria for morality become subjective: whether or not something is morally approvable or not depends solely on an agent's feeling of approbation or disapproval. FOOT 1978, 76–77, 78–79. To study how C.L. Stevenson, as a representative of emotivism, and R.M. Hare, as a prescriptivist, continue this tradition, see FOOT 1978, 96, 99. See also WALLACE 1986, 16–18.

that they describe the objects on which such evaluative statements are pronounced.[16]

Foot develops an alternative view for analyzing the relation between facts and values; according to her view, they are connected with each other in two ways. First, there exists a conceptual relation between facts and values; and second, factual premises can count as evidence in support of some evaluative conclusions.[17] A central part of emotivism and prescriptivism is the conviction that moral terms and those pertaining to facts belong to two strictly separable categories and that norms are connected with evaluation just because they contain an evaluative, or commendatory element. To think so is false, however, while this way of thinking presupposes that the link between facts and values is *external*. The externalist interpretation of the meaning of moral terms implies that assigning a moral term to something is not dependable on the quality, or nature of the thing, but derives solely from the fact that the utterer wishes to commend something and is ready to act in accordance with this commendation. If this were the true analysis, we could, Foot maintains, call actually anything good or bad by just saying that it contains an evaluative meaning for us, and that by this we commit ourselves to obeying an imperative that accords with our value-statement about the object in question.[18]

The correct way of analyzing the connection between facts and norms is to interpret their relation as *internal*. To call the relation internal is to maintain that every moral term includes certain criteria regulating the proper use and application of the term. Hence, when we attribute moral goodness or badness to something this thing is necessarily con-

---

[16] Foot 1978, 78–80. Wallace (1986, 16), and Geach (1977, 4–5) also maintain that moral judgements are not expressions of feelings but that they describe facts in the world.

[17] Foot 1978, 99. Foot does not explicitly define what the conceptual link between factual and evaluative statements is, but she is certain that there is this connection. She examines some instances of non-moral attribution as analogical cases; thus, one cannot hold just anything, say, dangerous, there must be some link between calling something dangerous and its being injurious. The same applies to moral terms and their use, moral qualities cannot be arbitrarily attributed to just anything. Foot 1978, 92, 106, 112–114.

nected with human good and harm. And attributing moral worth to something is not conditional on the speaker's willingness to commend the thing, but depends on what the thing itself is.[19]

Another serious flaw in mainstream ethical theories is the way these theories conceive the nature of morality as an institution. Virtue theorists hold that morality ought not to be seen individualistically, but in the context of a community. For Foot, the individualistic character of mainstream moral theories is a major flaw since there are many institutions, mores and practices which are pointless if we consider them outside a community in which such practices acquire their meaning, and in the life of which they are valid.[20] The whole validity of ethics does, in fact, reside in its being a social mutually supported institution.[21]

Foot's view here comes near ethical naturalism. She does not, however, presuppose that ethical terms depict reality in any objective sense in the meaning of ethical realism, she seems rather to represent a view akin to Putnam's internal realism. Morality is a mutually supported institution and the conventions of moral language create the reality within which moral terms receive a meaning and rules of correct use.[22]

Morality as a social institution also forms the correct background for

---

[18] FOOT 1978, 112. According to Foot's criticism, emotivism and prescriptivism are defenceless against moral "excentrics", who propose moral views that they have chosen to commend, but that other people would find weird as moral positions. Someone could, e.g., commend the clasping of hands as a good moral action, without any further explanation, just on the basis that she finds it commendable. Because of the possibility of such cases, it must be a mistake to place the connection of moral terms and evaluation in a specific evaluative element that is irrelevant in respect to what the thing itself is. FOOT 1978, 112–113. The main target of Foot's criticism is Richard Hare's analysis of moral terms and the moral theory based on the analysis, namely universal prescriptivism, see chapter "The logic of moral language" on page 50. For Hare's answer to this criticism, see HARE 1981, 63–64.

[19] FOOT 1978, 120. von Wright holds a position similar to Foot's, but he links his view to a conception of the status of morality; von Wright maintains that morality is not conceptually autonomous, but that moral good is only a secondary good, and it must be defined as an attribute of acts and intentions in terms of the *beneficial*. In this way, the meaning of the morally good derives from the good of man understood in a wider, non-moral sense. VON WRIGHT 1968, 17–18.

[20] FOOT 1978, 189–190.

[21] See FOOT 1978, 202, 203, 207.

examining moral goodness and virtue. The task of morality is to have an impact on people's actions in so far as these affect other people's well-being. In the social life of a community certain human features and inclinations are defects in relation to the good of the community and to human flourishing. If human nature were different, things that we now call virtues would not necessarily count as virtues in those other circumstances. Hence, the form and function of virtues depend on what human beings and the world they inhabit are like: virtues have a corrective role in human life.[23]

To explicate Foot's view a step further, we can say that human inclinations have, at least partly, an effect which is contrary to enterprises and practices which humans regard as good, and it is thence the value and necessity of virtues arise. Foot's view of the importance of virtues is teleological, but not in the classical sense. Virtues are valuable because they have certain effects regarded as valuable within and for a human community, but virtuous life is not, as it is in classical ethics of virtue, a necessary constituent of the universal human telos.[24] Consequently, virtues receive only an instrumental status in Foot's ethical thinking. A person acts virtuously either because of her psychological disposition, or because being virtuous has a connection to what the person wants irrespective of the moral point of view.[25] In this respect, Foot's moral psychology seems to be basically Humean.

Foot's instrumental view of the virtues relates to her conception of morality in connection to human action. Foot claims that morality does not provide us with any specific reasons for action but that the reasons for being moral and for taking ethical considerations seriously can and

---

[22] See, Hilary PUTNAM's, article "There is at least one *a priori* truth." in his *Realism and Reason: Philosophical Papers, vol III*, Cambridge 1983, 98–114.

[23] FOOT 1978, 3–4, 8–10, 153.

[24] Foot refers to the problem concerning the relationship between virtues and a person's whole life as she discusses the virtue of wisdom (FOOT 1978, 5–7) and the dilemma arising from virtuously performed immoral actions (FOOT 1978, 14–17). Foot comes to the conclusion that there should be some coherent "total" view for ordering the virtues in connection to a life, but she is not ready to adopt the view on human nature represented by the classical ethics of virtue, and she does not proceed to develop any such view herself.

must be equated with any other reason for intentional action; the acting person must believe that her action furthers her desires or accords with her interests.[26] Consequently, moral statements are not categorical, but only hypothetical imperatives. In this respect they do not deviate from any other, non-moral rules of a society, such as rules of etiquette.[27] The hypothetical character of morality in Foot's theory implies that morality cannot provide us with universally valid reasons for action: we have a reason for acting as far as the suggested action accords with our desires and/or interests.[28]

Wallace bases his ethical view on the assumption that the conception

---

[25] FOOT 1978, 127–131. Geach joins Foot in insisting that reasons for action, whether moral or not do not make sense unless they are connected with our inclinations. He simply maintains that we have a reason to act morally even if this is sometimes contrary to our own interest while we are naturally inclined to care what happens to others. GEACH 1977, 17. His discussion on the topic is short, and he presents his view in relation to his criticism against Kantian ethics of duty: we must connect our moral reasons for acting with our inclinations, otherwise we cannot say that acting morally were rational. GEACH 1977, 9.

[26] Foot maintains that even if it is correct to say that moral judgements stand for any person, it does not follow from this that an acknowledged moral demand establishes a reason for action for just anyone. Accordingly, it cannot be said to be irrational to be immoral. Moral judgements are connected with intelligible human pursuits and practices, but this does not , nevertheless, guarantee that everyone had a reason to choose the morally good. FOOT 1978, 152. See also FOOT 1978, 130–131. With her view Foot opposes Kant and his theory of the categorical imperative: she does not accept the view that moral considerations would necessarily present universal reasons for acting irrespective of people's individual desires and interests. FOOT 1978, 154, 162. For Kant, moral considerations are overriding because in his theory moral reasons must coincide with rationality. This is, according to Foot, an untenable position: there is no necessary connection between morality and rationality, there are cases in which it would benefit the person not to act morally. FOOT 1978, 152, 153. Kant was partly led to his mistaken view by his faulty theory of human nature. Kant was, according to Foot, actually a psychological hedonist who thought that ethics, if based on desires and interests, would necessarily be egoistic. FOOT 1978, 165.

[27] "The conclusion we should draw is that moral judgements have no better claim to be categorical imperatives than do statements about matters of etiquette. People may indeed follow either morality or etiquette without asking why they should do so, but equally well they may not." FOOT 1978, 164. "The exceptions to moral rules are built into the verdictive moral system and so it is *taught* that morality is always to be obeyed. With etiquette it is different, and therefore the rules of etiquette appear as a set of 'conditional' commands." FOOT 1978, 188.

[28] FOOT 1978, 130–131.

of *human life* is in itself normative in the sense that it incorporates the concept of human good. According to Wallace, it is impossible to talk about biological organisms and their lives without some notion of what their normal life and its sustenance, that is to say, their well-being, and good necessarily require.[29]

Wallace joins Foot's criticism of individualistic moral theories, but in his theory the social character of morality receives a more integrated position. According to Wallace, the first question we must ask in moral philosophy is "What is characteristic of human life?" For Wallace, human life and its normal progress presuppose that we have means for obtaining nutrition, and that an environment enabling growth and reproduction is available. These are the prerequisites humans share with other biological organisms, but what is uniquely human is "a life informed by convention".[30] *Conventional activities* include all endeavours which can be called *rule-following behaviour*.[31] The conventional aspect is central to all human life, and this is something innate to human beings, or a natural phenomenon. Wallace clarifies his point by saying that a human being

---

[29] "It must be conceded at the outset that the conception of human life used in the following discussions is a normative one. This usage, however, is unavoidable, because life is a normative concept that cannot be understood apart from the conception of a creature's good. It does not follow from this, however, that dicta about a certain creature and its life divide into objective facts and subjective normative notions that exist in an epistemological vacuum. Among the facts about living creatures are how they live normally, under what conditions they flourish or languish, and what the proper functioning is of their parts. Knowledge of such things is indispensable for the biological sciences. [...] There is no reason in principle why a study of human excellences based upon the nature of human life need be any less objective, well founded, or authoritative than the study of any sort of living creature." WALLACE 1986, 16–17. See also VON WRIGHT 1968, 105–112.

[30] WALLACE 1986, 34. Wallace applies Aristotle's scheme of human life and excellence in his theory, but instead of saying, like Aristotle, that the feature constituting a good human life is a life in accordance with *logos*, he maintains that a social life informed by convention characterizes what is distinctively human. Human excellences are, accordingly, tendencies and capacities for living well a social life informed by convention. WALLACE 1986, 37; 1988, 224–225.

[31] "I am using the term 'conventional', in a broad sense to include all of the things that philosophers have called institutions and practices. By conventional activities I understand all the activities that philosophers, influenced by Wittgenstein, have called "rule-following" behavior." WALLACE 1986, 34.

who is incapable of taking part in activities which require conventions is imperfect in a way comparable to an organic defect in any animal or plant.[32]

Seen from this perspective, we can say that conventional activities constitute the specifically human, or the *differentia specifica* of the human species in Wallace's theory. It is noteworthy that this characterization of what makes human beings what they are already contains a normative, evaluative aspect. Wallace regards the tendency in human behaviour to take a form of a convention as hereditary. He does not, however, make a distinction between this characteristic as a general human feature, and the specific form which these conventions take subject to the conditions of each particular community. This gives Wallace's theory a naturalist colouring and makes it vulnerable to certain kinds of criticism. The prevailing conventions of any community receive a status similar to that of natural facts. Their being what they are seems justified simply through their sheer existence. This makes Wallace's theory prey to cultural and moral rigidity and conservatism. Neither does the theory help to explain how and according to which criteria social conventions change or should be changed.

Wallace calls a community in which people share certain central crucial conventions and practices, a *form of life*.[33] A form of life constitutes the background against which the social conventions and practices of the community make sense.[34] It also serves as the context in which morality is rational and valid.[35] If we adopt this point of view it becomes superfluous to ask, why be moral, or why conform to moral rules. A

---

[32] WALLACE 1986, 34, 35, 110. Wallace's view allows a vast variety of realizations of the good human life, and it is perhaps not even possible to point out conventions that were absolutely beneficial to all human communities. Still, Wallace's view is not relativistic: some communities do further the well-being of their members better than others do theirs; and whatever variations there are in human communities, such things as health, intelligence, conscientiousness, benevolence, restraint, and courage foster the community, and their lack makes it worse off. WALLACE 1986, 35–36.

[33] WALLACE 1986, 112.

[34] WALLACE 1986, 104, 105.

[35] WALLACE 1986, 112, 113–116; 1988, 223.

community requires of its members, not just certain kinds of acts, but certain ways of behaviour; otherwise it would not be a community. The members of the community comply to such ways of behaviour, or in other words, they act morally, because for them it seems reasonable that such behaviour is required of them as members of the community. People also observe the requirements of morality despite the fact that being moral is sometimes contrary to their individual desires and interests.[36] Wallace maintains that from this point of view it is actually a severe mistake to consider reasons for acting morally in terms of personal interest and desire. Morality and its validity do not rest on the assumption that they accord with individual interests and desires, and people's motivation to act morally is not based on their will to use morality as an instrument for gaining what they want.[37] Morality is a social phenomenon constituted and sustained by complex conventions and the existence of these conventions is enough to constitute both a reason for acting and a motivation to act morally.[38]

To explicate Wallace's thought further, we can say that morality makes a group of people living together a community, and similarly, that human beings become responsible members of their community, that is, moral persons, through morality. There also prevails an analogy between a con-

---

[36] Wallace 1986, 115–116.

[37] Wallace criticizes utilitarian and Humean types of ethics for placing the justification of moral action and reasons for acting in the wrong place: "There is a tendency to take as *the* paradigm of practical reasoning an individual's trying to figure out how to get what he wants. Good practical reasoning is conceived exclusively as discerning the most effective and efficient means to the desired end. The rationality of pursuing the end may come into question by asking how its pursuit may affect one's chances of getting other things that one wants, but the pattern here is fundamentally the same. The reasonableness of one's pursuit of a particular end is assessed in terms of its conduciveness to other particular ends one desires. If one desires something for its own sake, however, then this desire itself is neither reasonable nor unreasonable." Wallace 1986, 113. This form of argument proves to be especially problematic when we have to explain why somebody not only chooses an act contrary to her interest, but when somebody is committed to a certain form of behaviour, like a conscientious person is, and this commitment does not operate as a means of achieving something else. Referring to the self-evidence of duties supporting such forms of behaviour cannot serve as a model of justification for morality, as for example Prichard suggests. Wallace 1986, 114.

[38] Wallace 1986, 117–118.

ventional form of life in a community and an individual's life. Conventions direct the life in a community in the same way as virtues direct a person's life and action. Good conventions make social life good, and virtues have the same effect on an individual life. A life in a society ordered by conventions cannot be analyzed into separate happenings, and the same applies to a good human life: it is not a collection of distinct actions but such life is an expression of a well-functioning character, and realization of virtues,[39] which are traits that enable and further the kind of life that is characteristic of human beings.[40]

Wallace comes near the classical virtue tradition in his view of the meaning of virtues. In addition to having a function as instruments for realizing central communal goods, virtues and the conventions supported by them are constitutive of a good human life. From this angle, virtues are both means to something else, and ends in themselves as far as good living is concerned.[41]

To conclude the present discussion, we can say that both Wallace and

[39] WALLACE 1986, 10. WALLACE 1986, 153. A good character is of crucial importance for moral goodness, for a firm determination to do what one morally should will not necessarily, and not even usually lead to the morally desired action. A morally good agent takes the complexity of the situation into account, and this requires a complex and sensitive moral character. WALLACE 1986, 126. Certain central virtues, such as restraint, benevolence, and justice, are not optional for the well-being of a community in the way, for example, cleanliness or amiability are, but these virtues are necessary for a variety of human goodness in such a way that their removal from our lives would endanger the whole structure. No community, and not even rationality are possible if the members of a community do not practice these virtues, at least to some extent. Wallace is even ready to maintain, that: "[i]f the virtues were removed from the scene, however, human life — the form of life characteristic of our kind — would be impossible." WALLACE 1986, 153. Even if the virtues receive a central place in Wallace's ethical theory they do not replace moral rules: for Wallace, moral rules have an indispensable role in every moral theory. This is due to the fact that some of the most central virtues concern the person's attitude and observance of moral rules. WALLACE 1986, 9.

[40] WALLACE 1986, 36–38, 161. WALLACE (1988, 230–231) stresses that according to his view of virtue ethics, we do not have to presuppose a *telos* common to all human beings. Instead of seeing the virtues as directed towards realizing a "fixed goal" inherent to human nature, we should understand them as being analogous to practical crafts. Thus, "just as the exercise of the skills and capacities of the good carpenter constitutes the activity of practicing the craft well, so the exercise of the virtues constitutes the activity of living well." See also WALLACE 1991, 178–179.

[41] WALLACE 1986, 156, 159–161.

Foot regard morality as a social institution. They both stress that morality is not subjectivist, depending on individual sentiments but that moral terms do describe real non-subjective properties in a way that accords with how things are understood in the community. Despite this similarity, there is a difference of emphasis between Foot and Wallace. Wallace regards the topic from what could be called a communitarian viewpoint, whereas Foot examines the issue from the perspective of the philosopy of moral language. Furthermore, Wallace locates human action within the context of conventional activities, whereas Foot joins the Humean tradition of explaining human behaviour. Wallace follows the traditional ethics of virtue in his understanding of virtues and conventions sustained by virtues as constitutive of good human life, but Foot adopts the virtues more as an instrument for realizing some individually or communally desired state of affairs. Against the background of utilitarianism and deontological theories the present approach makes the problem concerning the justifiability of morality as an institution less problematic. Moral activity is meaningful and rational within the context of a community. Consequently, the moral person is placed in a new setting. Morality is no longer something external to what the person is, but can be conceived as establishing a point of view which constitutes the agent's understanding of herself as a moral person.

3.2. A new approach to ethics — Nussbaum, MacIntyre and Taylor

As different as the works of Martha C. Nussbaum[1], Alasdair MacIntyre[2] and Charles Taylor[3] are, they do have three features in common.[4] First, they all heavily criticize the dominant trends of modern moral philosophy, which especially MacIntyre and Taylor see as reflecting some more general tendencies of Western thinking. The second common feature is their interest in historical material. Although they do not regard themselves as simply historians they all examine historical texts and historical developments in an unprecedented way compared to the thinkers we have considered so far. They all claim to take part in the present discussion of moral philosophy, but their use of historical material suggests an approach essentially deviating from that of, e.g., Foot's and Wallace's. The third factor connecting the work of Nussbaum, MacIntyre and Taylor is their attempt to create a new theoretical approach to ethical questions. Here their criticism of prevailing moral philosophy reveals the need for an alternative approach. The historical material is then applied to the search for a new ethical view. But what is the outcome? Can we find a fresh theoretical, or meta-ethical approach to moral philosophy? If there is a new theoretical setting involved does this have an impact on

---

[1] Martha NUSSBAUM, *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge University Press, Cambridge, 1986; *Love's Knowledge. Essays in Philosophy and Literature*. Oxford University Press, New York and Oxford, 1990.

[2] Alasdair MACINTYRE, *After Virtue: A Study in Moral Theory*. Duckworth, London, Second edition, 1987; *Whose Justice? Which Rationality?* Duckworth, London, 1988; *Three Rival Versions of Moral Enquiry*. Duckworth, London, 1990.

[3] Charles TAYLOR, *Sources of the Self: Making of the Modern Identity*. Cambridge University Press, Cambridge, 1989; *Human Agency and Language. Philosophical papers 1*. Cambridge University Press, Cambridge, 1990; *Philosophy and the Human Sciences. Philosophical papers 2*. Cambridge University Press, Cambridge 1990.

[4] NUSSBAUM (1992, 10) emphatically states that although both she and MacIntyre represent the newly revived ethics of virtue their views are actually very far from each other, the main difference being, according to Nussbaum's interpretation, the meaning of universality and the role of reflection in an ethical theory.

the concept of a person that these theories represent? It will be as well to keep these questions in mind as I continue my analysis of virtue theories. In the following, I will first examine Nussbaum's theory and then move on to consider the contributions made by MacIntyre and Taylor.

3.3. Nussbaum's *perceptive equilibrium*

Nussbaum characterizes herself as an Aristotelian ethicist. Her aim is to show that both utilitarian and deontological moral theories represent a one-sided, and thus defective view of ethics. There are several features which have long been neglected in modern moral philosophy but which play a central role in a comprehensive understanding of human life. To correct the flaws of prevailing moral philosophy Nussbaum introduces a revived Aristotelian approach and a corresponding method, *perceptive equilibrium,* to ethical thinking. In the following, I will examine Nussbaum's criticism of rival moral theories, and analyze the perceptive equilibrium as a proposed corrective for doing ethics. I will then make use of this analysis to explicate the concept of a person entailed by Nussbaum's views.

Nussbaum criticizes several features of modern moral philosophy, most of them shared by both utilitarian and Kantian or deontological models.[5] First, the dominant moral theories treat ethical values as something *sui generis*, as separable and distinct from other kinds of values. Connected with this view, these theories also interpret ethical values as

[5] Nussbaum classifies John Rawls as Kantian, although she sees him as one of the endorsers and users of an Aristotelian ethical procedure the end of which is practical, and socially practicable. NUSSBAUM 1990, 25, 173, 185. Rawls' mistake is that he, unlike Aristotle, and Nussbaum herself, prizes generality and universality at the cost of particularity, and that, according to his theory, there is a general ordering that can be imposed upon all conflicting claims. NUSSBAUM 1990, 175. See below, (page 190) the discussion about the value of the general and the particular in ethics.

commensurable with each other in a way that makes them subtractable under one single moral rule, principle, or good.[6] Nussbaum disagrees with this view, arguing that a distinction made between different kinds of values can only be artificial. Since all values promote human good in their own characteristic way.[7] Moreover, various goods of human life are incommensurable with each other. It is from a complex plurality of values that the good human life emerges. Different values are non-exchangeable in a way that leads both to tragic conflicts and to moral dilemmas.[8] The incommensurability and plurality of ethical values imply that there cannot be any algorithm, or mechanical procedure for solving them.[9]

Another typically modern feature is that moral philosophy concentrates on the general features of human life. Modern moral theories regard moral rules and principles as explicators of general characteristics of each deliberative situation, and these rules receive a position of ultimate authority for determining the correctness of particular decisions.

[6] This claim applied to Kantian theories, see NUSSBAUM 1986, 4,5; and to Utilitarian theories, see NUSSBAUM 1986, 112–113; and to both of them, see NUSSBAUM 1990, 173. STOCKER (1990, 124–126) maintains that we cannot eliminate conflicts from our ethical thinking, no matter which kind of ethical theory we apply, because we live in a contingent world. This does not, however, have to threaten the possiblity of constructing a practicable, complete and realistic moral theory.

[7] In discussing the differences between the ancient Greek way of understanding the problems of practical rationality contrasted with that of Kantian theories, Nussbaum writes: "I shall, in fact, try to avoid not only the Kantian moral/non-moral distinction, but all versions of that distinction and of the related distinctions between moral and non-moral practical reasoning, moral and non-moral practical conflict. The Greek texts make no such distinction. They begin from the general question, 'How should we live?' and consider the claim of all human values to be constituent parts of the good life; they do not assume that there is any one group that has even a *prima facie* claim to be supreme. I believe that their approach is faithful to the way that our intuitive practical reasoning does in fact proceed, and that it recaptures aspects of our practical lives that tend to be obscured in works beginning from that distinction, however understood." NUSSBAUM 1986, 5.

[8] NUSSBAUM 1986, 7, 27–28; 1990, 36–37; 1992, 10. J.-P. Sartre's model of freely improvising our choices without any further regret; R.M. Hare's insistence on a logically coherent and non-conflicting set of principles as a prerequisite for adequate moral reasoning; as well as Kant's claim that it is part of the very notion of a moral rule or principle that it can never conflict with another moral rule, all exhibit a false quest for moral commensurability. NUSSBAUM 1986, 31.

As a result, these theories tend to neglect the particular aspects of the human situation. This is all the more serious a tendency because it appears together with an insistence on the commensurability of all ethical values; it involves a proclivity to impoverish our view of both life and moral philosophy. Consequently, the scope of ethical theory becomes narrow: only questions which can be formulated in terms of a fixed theoretical framework receive moral attention.[10]

Nussbaum makes a distinction between something being general and its being universal. She especially directs her criticism towards an improper understanding of general rules in moral reflection, not so much against the search for the universal in ethically salient situations. Concentration on the general features of a situation implies that one only pays notice to some characteristics of the case; and, further, that one treats them as independent of each other. Using Aristotle's examples, Nussbaum equates ethical reflection with practical skills like navigation and medicine, in the practice of which mechanical application of general rules would seldom lead to the desired aim. If attention is paid

---

[9] "We have seen, [...], that a contingent conflict between two ethical claims need not be taken for a logical contradiction; and that the 'inconsistency' between freedom and necessity can, similarly, be seen as a correct description of the way in which natural circumstances restrict the possibilities for choice." NUSSBAUM 1986, 46. About the incommensurability of human goods, see also NUSSBAUM 1986, 27–28. Nussbaum criticizes moral philosophers who insist that "in every case there is at most a single correct answer, and the competing candidate makes no further claim once the choice is made". NUSSBAUM 1986, 30. GRIFFIN (1990, 75–92, 333–334) distinguishes several forms of incommensurability including incomparability and discontinuity between different kinds of values. According to his view, even if we acknowledge the plurality of values in the sense that we deny the possibility of one super-value to which other values could be reduced, we can still construct a method for measuring different types of values against each other. To accept, say, money as a common measure is not to make it the supreme value.

[10] Nussbaum classifies Plato as the initial protagonist of this view: "As the aspiring Platonic philosopher scrutinizes the particular to see the universal features it exemplifies, and considers it ethically relevant only insofar as it falls under the general form, so the aspiring person of practical wisdom will seek to bring the new case under a rule, regarding its concrete features as ethically salient only insofar as they are instances of the universal. The idiosyncratic cannot be relevant. The universal principle, furthermore, is normative because of itself (or because of its relation to higher principles), not because of its relation to particular judgments." NUSSBAUM 1986, 299–300. See NUSSBAUM 1990, 23–24 about a similar tendency in Kantian and utilitarian theories.

solely to what is general the importance of the particular features of a situation are ignored. This mistake can be avoided if a given situation is examined in detail, in its actual context. The analysis will then reveal which features in that specific situation are morally relevant and thus universalizable. Furthermore, practice and experience in the deliberative activity will enable a person to develop their ethical decision-making.[11]

It is worth noticing in this connection that Nussbaum also gives ethical weight to questions concerning style. According to her view, there are many ethical topics which can be given a proper treatment only if they are presented in the form of novels and other typical literary genres. Contrary to the view of most theorists, Nussbaum values style as an essential part of any ethical project. This perspective reveals a further weakness in utilitarian and deontological theories as they display a tendency to focus on the general in moral deliberation. Any issue which cannot be formulated within the framework of a particular theory tends to fall outside the scope of proper or legitimate ethical interest. Such treatment of human dilemmas, however, contradicts our common intuition of life and its complexity and impoverishes our ethical project.[12]

The third target of Nussbaum's criticism is the role given to the intel-

---

[11] NUSSBAUM 1990, 38–40; 1992,.

[12] NUSSBAUM 1990, 5, 46–47. The aspect of style has almost completely been neglected in contemporary Anglo-American philosophy. NUSSBAUM 1990, 8; see, however, MURDOCH 1967, 1–2, 15. for a different view. In discussing the questions of ethical interest and importance presented in literature, and in reflecting what an appropriate theoretical framework for their study would be, Nussbaum writes: "A difficulty here is that some influential accounts of what moral philosophy includes are cast in the terms of one or another of the competing ethical conception; thus they will prove unsuitable, if we want to organize a fair comparison among them. For example, if we begin with the Utilitarian's organizing question, "How can one maximize utility?," we accept, already, a certain characterization of what is salient in the subject matter of ethics, of the right or relevant descriptions for practical situations — one that would rule out from the start, as irrelevant, much of what the novels present as highly relevant. Similarly, reliance on a Kantian characterization of the domain of the moral, and of its relation to what happens in the empirical realm, together with reliance on the Kantian's organizing question "What is my moral duty?," would have the effect of artificially cutting off from the inquiry some elements of life that the novels show as important and link to others — all in advance of a sensitive study of the sense of life that the novels themselves have to offer." NUSSBAUM 1990, 24.

lect on the one hand, and to the emotions on the other hand in ethical reflection. Dominant moral theories prize intellect as the only useful instrument in the process of deliberation. Consequently, emotions are either ignored as completely unimportant for ethics or as misleading and by their nature unreliable. Thus, the intellectualist, emotionally non-provoking approach is the adequate mode for presenting moral problems.[13] In Nussbaum's view, such a conception of the ethically relevant depletes our view of practical wisdom and vitiates the picture of human agency as well as that of the human condition.[14] A richer way of understanding the nature of ethical deliberation is to see it as involving both a person's intellectual reflection and her emotions. Emotions always involve a cognitive component and they are, in this sense, capable of contributing to rational ethical reflection. Intellectual reflection, again, is needed for evaluating not only distinct situations but also for weighing the appropriateness of conventional moral conventions and principles. Reflection is, moreover, necessary for developing a person's passions to accord with moral virtues contributing to and constituting a good life.[15]

The two points of Nussbaum's criticism I have discussed above, namely the tendency for moral theorizing to concentrate on the general characteristics and ignore the particular in human life, and the disregard of the emotional side of human existence, both focus on the same issue:

[13] "According to one version of the objection, emotions are unreliable and distracting because they have nothing to do with cognition at all. According to the second objection, they have a great deal to do with cognition, but they embody a view of the world that is in fact false." Nussbaum 1990, 40. See also Nussbaum 1990, 8. Here, too, we are dealing with an ancient quarrel, that is, the dispute concerning the cognitive value of emotions in ethical deliberation. Nussbaum dismisses the view that emotions are blind animal reactions, that they have nothing to do with cognition, as uninteresting, as one containing too impoverished a conception of emotion, to survive any closer scrutiny. Nussbaum 1990, 40–41. In contrast, the second view, i.e., that emotions are misleading guides in human deliberation, has received much support during the history of Western thought. Nussbaum 1990, 42. For a view similar to Nussbaum's, see Murdoch 1967, 6–7.

[14] Nussbaum 1986, 186. Emotions do not only play a cognitive part in ethical deliberation, there are also aspects, and values in life which one will not be able to apprehend without the help of certain emotions adequate for the particular situation. Human life will be decisively more meagre if the impact of emotions in our moral life is depleted. Nussbaum 1990, 40–42. See also Nussbaum 1986, 214–216; 307.

[15] Nussbaum 1992, 11.

prevailing moral theories make our view of morality too tight. Nussbaum wants to avoid defining morality in any specific and general way in order to abstain from excluding any aspect of human life from ethical consideration. But does Nussbaum's view imply that all aspects of human life get transformed into moral questions? Do we not lose all our criteria for distinguishing the moral from the non-moral, and does not everything thereby become *a priori* morally relevant? Nussbaum's view then appears to be clearly different from the approach we encountered as we examined theories from the social contract tradition: their salient feature was the clear distinction between the moral and the non-moral. We cannot, however, yet say whether Nussbaum's wider view of the ethically relevant comes close to the utilitarian tendency towards "moral imperialism", the view that everything in human life has moral significance. I will return to this question later on.

The remedy for the faults of the current ethical theories is a return to Aristotelian ethics. This does not, however, mean that we should adopt Aristotelian ethics in its substantive form, but that we should, instead, start employing the Aristotelian approach and method in our moral thinking. Nussbaum maintains that she, in her own writing, exercises the Aristotelian method. This involves making use of Aristotle's texts in two respects. First, Nussbaum sees them as texts which introduce and apply a specific method in ethical reflection. By studying Aristotle's texts from this point of view one can learn his method. Second, Nussbaum treats Aristotle's writings just the seme as any other texts relevant to the subject matter; she examines them as arguments, suggestions, or comments in the discourse that forms the material for the Aristotelian dialectic or the Aristotelian procedure in ethical study.[16]

The fact that Nussbaum claims the label Aristotelian for herself also explains her use of historical material. She treats all texts, whether ancient or modern, as material displaying human life in its richness and variety. This means that her project is not primarily that of writing historical treatises on moral philosophy; in works on Greek thought she takes part in an ongoing ethical discussion, in which the reading and interpretation of ancient texts form an important part.[17] In addition, Aristotle's

texts receive a special position in Nussbaum's project because they not only provide material for ethical study but also offer the method for doing moral philosophy.[18]

Nussbaum's aim is to introduce an extensive ethical method that not only covers the same areas of moral phenomena as, say, utilitarian and Kantian theories but that also corrects their mistakes. Nussbaum calls her version of the Aristotelian inquiry *perceptive equilibrium*.[19] The procedure starts with the question: "How should a human being live?".[20] To explicate what is involved here, we can say that the aspect on which the question focuses is human life as chosen, and not as given. In other words, life is not looked upon from a third-person perspective, as a causally determined sequent of events, but from a first-person point of view,

[16] "Here [in trying to find a proper starting point for ethical reflection] both life and the history of philosophy combine to help us. [...] And in the history of moral philosophy we also find an account of an inclusive starting point, and an open and dialectical method, that is, in effect, the philosophical description of this real-life activity and how it goes, when done with thoroughness and sensitivity. For the proponents of rival philosophical conceptions in ethics have usually not concluded that their inquiries and results were non-comparable with those of their opponents, or comparable only by a method of comparison that already throws the judgment to one or another side. They have, instead, frequently appealed to the inclusive dialectical method first described by Aristotle, as one that (continuous with the active searching of life) can provide an overarching or framing procedure in which alternative views might be duly compared, with respect for each, as well as for the evolving sense of life to which each is a response. [...] I [...] follow this example — insisting, as well, that one of the salient virtues of this method is its continuity with "our actual adventure" as we search for understanding. (It is important to distinguish the Aristotelian procedure and starting point from Aristotle's own ethical conception, which is just one of the conceptions it considers)." Nussbaum 1990, 25. See also Nussbaum 1990, 55; 1992, 11.

[17] After describing the ancient Greek discussion on effects of luck on moral goodness, and its different features (Nussbaum 1990, 16–18), Nussbaum writes: "But my interest in the ancient debate was motivated by an interest in philosophical problems whose force I felt, and feel, in life. I therefore began, while pursuing these historical issues, to look for ways of continuing the ancient debate in the contemporary philosophical context. The ancient debate had helped me to articulate much that I had sensed long before about the novels of Dickens and Dostoyevsky [...]" Nussbaum 1990, 18.

[18] Nussbaum 1990, 24–25.

[19] Nussbaum 1990, 182–183, 187.

[20] "The Aristotelian procedure in ethics begins with a very broad and inclusive question: "How should a human being live?" Nussbaum 1990, 25. See also Nussbaum 1990, 173; 1992, 10.

as a set of chosen possibilities. Nussbaum admits that if we base our ethical inquiry on this question we actually commit ourselves to a certain point of view. Doing this is not fatal for the project, however, for there are no completely neutral positions; the Aristotelian alternative, as well as any other approach, already suggests a set of possible answers. The fact that we must adopt something as a starting point is a threat to ethical inquiry, for there is always a danger of being biased, subjective, or irrational.[21] Despite the risk of partiality, we must take up the procedure, since giving up asking basic questions about human life equals with abstaining from pursuing practical aims with intellectual activity.[22]

The form of the basic question, "How should a human being live?", as well as the name of Nussbaum's project, perceptive equilibrium, both refer to the twofold objective of the Aristotelian inquiry, which is *empirical* and *practical*. Its being empirical, and therefore perceptive, means that the experience of life supplies the evidence and material for the procedure. The other aim of the procedure is practical, involving the search for an equilibrium, the goal of which is the formulation of a conception by which human beings can live.[23] I will first examine the empirical feature in Nussbaum's procedure and thereafter scrutinize what its practical nature implies.

To minimize a possible bias in one direction or another, our ethical

[21] "No starting point is altogether neutral here. No way of pursuing the search, putting the question, fails to contain some hint as to where the answers might lie. Questions see things up one way or another, tell us what to include, what to look for. Any procedure implies some conception or conceptions of how we come to know, which parts of ourselves we can trust. This does not mean that all choices of procedure and starting-point are merely subjective and irrational. It does mean that in order to attain to the rationality that is available (as the chimera of total detachment is not) we need to be alert to those aspects of a procedure that might bias it unduly in one direction or another, and to commit ourselves to the serious investigation of alternative positions." NUSSBAUM 1990, 24–25.

[22] "What I propose here is not a merely theoretical undertaking, but one that is urgently practical, one that we conduct every day, and must conduct. If we wish to regard the obstacles against fair comparison of alternatives as insuperable for reasons of methodological purity, we can always do so; but at enormous practical cost. And our common experience, our active practical questions, give a unity and focus to the search that it might not seem to have when we regard it solely on the plane of theory." NUSSBAUM 1990, 28.

approach must be empirical. For the same reason, it is also necessary not to limit the object of ethical study and not to restrict the variety of possible empirical answers. A strictly theoretical approach will easily lead to the exclusion of some features of human life from scrutiny. In accordance with these "precautions" against any bias caused by hastily adopted presuppositions, Nussbaum's Aristotelian method does not allow a categorical division of values filed under "moral" and "nonmoral". Our view of actual human life from which the ethical procedure must derive its material must, empirically speaking, be as inclusive as possible.[24] A reason in favour of this approach is that it grants a fair treatment to its rivals, too. The utilitarian and the Kantian models for a good human life represent, namely, two major alternative answers to the initial question of the ethical enquiry.[25]

Nussbaum defines her ethical procedure as Aristotelian but it may, as a matter of fact, violate a feature central to Aristotle's own method. In Aristotle's view, all entities, human beings included, have their own proper place in the structure of the universe. Now, to live a good human life necessarily requires taking this structure into account. As far as ethical inquiry is concerned, this implies that not just *any* alternative will do as a tentative answer for the good of human life, but only options that

---

[23] "The inquiry [...] is both empirical and practical: empirical, in that it is concerned with, takes its "evidence" from, the experience of life; practical, in that its aim is to find a conception by which human beings can live, and live together." Nussbaum 1990, 25. "The inquiry asks, then, what it is for a human being to live well. This investigation as I imagine it, is both empirical and practical. Empirical in that it is based on and responsible to actual human experience [...] Practical in that it is conducted by people who are themselves involved in acting and choosing and who see the inquiry as having a bearing on their own practical ends." Nussbaum 1990, 173.

[24] "The question [how should a human being live] presupposes no specific demarcation of the terrain of human life, and so, *a fortiori*, not its demarcation into separate moral and nonmoral realms. It does not, that is, assume that there is, among the many ends and activities that human beings cherish and pursue, some one domain, the domain of moral value, that is of special importance and dignity, apart from the rest of life. Nor does it assume, as do utility theorists, that there is a more or less unitary something that a good agent can be seen as maximizing in every act of choice. It does not assume the denial of these claims either; it holds them open for inquiry within the procedure — with the result that, so far, we are surveying everything that Aristotle surveys, that we do actually survey: humor alongside justice, grace in addition to courage." Nussbaum 1990, 25.

accord with the structure of things.[26] Nussbaum's ethical project seems to depart from the Aristotelian in this strict sense.

The name Nussbaum has given to her ethical procedure also reflects her insistence on the inclusiveness of the ethical point of view; the procedure is *perceptive*. Nussbaum stresses that the material used in ethical reflection has to be as vast and rich as life itself. Consequently, importance must be attached to the source of the material and the way in which it is collected. The purpose of this "wide perceptiveness" is to prevent imposing our presuppositions and ready-made categories onto normative questions about human life.[27] The study must be comprehensive, including all aspects of human life, not only the general, the universal, the public, and the non-controversial, but also particular features, emotional aspects, as well as the ambiguous and messy parts of human existence.[28]

What does ethical inquiry then involve? In Nussbaum's view, we must try to find an answer to the basic question "How should a human being live?" in the sense of "What are the most common problems of human

---

[25] "I have said that the question with which my projected literary-ethical inquiry begins is the question, "How should one live?" This choice of starting point is significant. This question does not (like the Kantian question, "What is my moral duty?") assume that there is a sphere of "moral" values that can be separated off from all the other practical values that figure in a human life. Nor does it assume, as does the utilitarian's focus on the question, "How shall I maximize utility?" that the value of all choices and actions is to be assessed in terms of a certain sort of consequence that they tend to promote. It does not assume the denial of these claims either. So far it is neutral, leaving them for investigation inside the inquiry. The point is to state the opening question in a general and inclusive way, excluding at the start no major story about the good life for human beings." NUSSBAUM 1990, 173.

[26] See, *Met* XII, 7; and *EN* X, 7–8. C. C. W. TAYLOR (*Mind*, 1987, 411) offers an analogical objection to Nussbaum's way of interpreting Aristotle's conception of ethical truth. Taylor criticizes Nussbaum for introducing Aristotle as a supporter of some version of the coherence theory of truth, whereas Aristotle in reality represents a very firm version of the correspondence theory of truth. But there is no evidence for the claim that Aristotle has a special conception of ethical truth apart from the one he maintains to hold for the rest of reality.

[27] It is possible that we, by following the Aristotelian procedure, end up with something that is alien to certain theoretical alternatives, but this is, according to Nussbaum, not a fault of our procedure, but derives from those other theoretical alternatives. NUSSBAUM 1990, 27.

life and what is a good way to face them?". This means reflecting on such universally human issues as the temporal nature of human existence, material need and the scarcity of resources as well as the problems concerning their distribution.[29] The search for an answer to these questions must be "perceptive", taking seriously the empirical material offered by life in its complexity. This does not, however, mean that we simply have to observe the crude, uninterpreted phenomena of life; instead different kinds of texts, historical and modern, philosophical and literary, scientific as well as artistic offer the material for the inquiry. These texts represent varying answers to the basic question. The stylistic variety of the material corresponds with the diversity of our cognitive faculties, with the "how" of ethical perception: there are aspects in human life which can only be expressed in a form and with a style of narration.[30] We perceive these features in their enormous richness, not with our intellect alone, but by feeling and sensing them, through our emotions. The goal we as perceivers must aim at is to become "finely aware" of all aspects of human life, taking account of the universal and general requirements of ethical norms, as well as the changing and unique aspects of the particular human situation.[31]

---

[28] In this connection Nussbaum criticized John RAWLS' *A Theory of Justice* (Clarendon Press. Oxford, 1972), and the way Rawls applies the Aristotelian method in his theory. Rawls calls his procedure *reflective equilibrium*, from which Nussbaum distinguishes her own mode of inquiry by calling it *perceptive equilibrium*. The weakness of Rawls' theory is, according to Nussbaum, that he restricts and limits the form of acceptable judgements at the initial stage of the procedure, and sets strict criteria for the final form of the theory, so that the theory disregards several aspects of human existence. Rawls ostracizes judgements that do not meet the criteria he sets for "considered judgment", i.e., conclusions that are made with hesitation, or without firm confidence, or, again, in a state of emotional turmoil. Rawls further formulates five additional "constraints that must be met by any ethical theory that will even be seriously considered during the procedure of scrutiny. These conditions are that its principles should be *general* in form and *universal* in application; that they should be *public* and available to all, that they should impose a general ordering on conflicting claims, and that these principles should be regarded as final and conlusive[...]" NUSSBAUM 1990, 174–175; RAWLS 1972, 47–48, 130, 135. For Nussbaum as well, one of the tasks of an ethical procedure is to question such basic requirements, and consider everything, including theories that abandon the criteria of methodological acceptability, so central to Rawls' theory, see page 110.

[29] NUSSBAUM1992, 11.

One could claim that the aims Nussbaum has established for her "perceptive" theory are contradictory. How can it be that the study of human life must be maximally extensive, including all possible aspects of life, and that it must at the same time find an answer to the substantive question "How should a human being live?" Does not answering this question necessarily mean that we have to choose something particular and exclude something else? And does not already accepting this question determine the answers we can find for it? Nussbaum is aware of this difficulty but sees it as no obstacle to her theory. First, the procedure is open-ended in the deep sense that it does not determine even such central issues as the norms of good or the criteria for rationality. The only requirement is that to accept something necessarily rules out its negation as a possible choice; nothing else is fixed.[32] Second, the procedure leads somewhere. It is not just an irrelevant enumeration of different alternatives, but both the starting point of the procedure and its outcome have relevance for people's lives. Although one has to consider a variety of

---

[30] Nussbaum insists that the material of ethical study, i.e., the empirical, must, besides the ordinary philosophical material, include novels and other literary texts, because "[b]uilt into the very structure of a novel is a certain conception of what matters", and this conception differs from the notions of most philosophical texts. NUSSBAUM 1990, 26. So: "The proposal is that we should *add* the study of certain novels to the study of these [philosophical] works, on the grounds that without them we will not have a fully adequate statement of a powerful ethical conception, one that we ought to investigate." NUSSBAUM 1990, 27. One of Nussbaum's main arguments is that style is not an arbitrary detail in any writing, but that it is always connected with the subject matter of the theme, it includes a commitment to certain values and points of view, to what is regarded as mattering. NUSSBAUM 1990, 18–19, 22, 26. There are invaluable areas of human life and experience, that cannot be adequately described by a philosophical, or scientific style. If we want to capture these sides of the human project, we must turn to literature. NUSSBAUM 1990, 281.

[31] Nussbaum equates the task of a skilful moral agent with that of an artist: "So, if we think of the perception as a created work of art, we must at the same time remember that artists, [...], are not free simply to create anything they like. Their obligation is to render reality, precisely and faithfully; in this task they are very much assisted by general principles and by the habits and attachments that are their internalization." NUSSBAUM 1990, 155.

[32] "Nothing is held unrevisable in this process, except the very basic logical idea that statement implies negation, that to assert something is to rule out something else." NUSSBAUM 1990, 26. See also NUSSBAUM 1990, 174; and NUSSBAUM 1986, 247, 252–253.

particular alternatives in one's ethical reflection, the initial question is universally human. This means that when an answer is found it is meaningful for human beings in general. Furthermore, after all possibilities have been weighed something is chosen that is practicable and shareable.[33] This choice is not, however, absolute or "eternal" but the method involves a readiness to change the answers in accordance with the challenges life imposes.[34] The required readiness to change the adopted position leads to the second central feature in Nussbaum's ethical method, namely, to its nature as an equilibrium.

Nussbaum characterizes her method by calling it an *equilibrium*. The name refers to two features in the theory; first to the treatment of the material under study, and second, to the nature of the practical aim in the procedure. In *perceptive equilibrium*:

> The central procedural idea is that we work through the major alternative views about the good life, holding them up, in each case, against our own experience and our intuitions. The first step will be to get a perspicuous description of these alternatives [...] Prominent among these views will be views embodied in texts of many kinds, both recent and older. Next we notice and clearly describe the conflicts and tensions among the views that we find. Where there is inconsistency or irreconcilable tension — and where this tension corresponds to something that we notice in our own experience and thought (individually or communally) — we aim to revise the overall picture so as to bring it to harmony with itself,

[33] Nussbaum 1992, 11.

[34] Nussbaum 1990, 27, 173, 174. Nussbaum admits that, no matter how open-ended the procedure is, there is still always a commitment to certain norms of procedural rationality embedded in the substantive conclusions they support. We cannot escape this problem just by yielding to the claim that all traditions and their adapted criteria of rationality are incompatible with each other: we always compare and seek the best way to live, there is no escape from that. Still: "The Aristotelian procedure tells us to be respectful of difference; but it also instructs us to look for a consistent and sharable answer to the "how to live" question, one that will capture what is deepest and most basic, even though it will, of necessity, to achieve that aim, have to give up certain other things. To this extent its flexibility is qualified by a deep commitment to getting somewhere. It is built into the procedure itself that we will not simply stop with an enumeration of differences and with the verdict that we cannot fairly compare, cannot rationally decide." Nussbaum 1990, 28.

> preserving, as Aristotle says, "the greatest number and the most basic" of the original judgments and perceptions. There is no rule about how to do this.[35]

The aim of perceptive equilibrium is to find a concept of human life which is internally coherent, accords with our intuitive sense of life and is broadly shared and sharable.[36] This project never ends: the perfect equilibrium is never attained for perception always offers new material for the procedure, and hence, there has to be a readiness to reconstitute the equilibrium in response to the new.[37] This does not, however, mean that the answers this sort of ethical reflection gives were relativist or purely particular. They have general meaningfulness and universal significance in that the questions to which they are answers are universal and central in the lives of all human beings.[38]

Nussbaum claims that her perceptive equilibrium is, in comparison to deontological and utilitarian models, a better ethical method. For this reason we should adopt it as a basis for our ethical thinking. This does not mean, however, that these other theories will have to be ignored and forgotten. Instead, Nussbaum promises a fair treatment for them within the perceptive equilibrium itself. Utilitarian and deontological theories form part of the perceived material in ethical study. But what does this mean? Does Nussbaum maintain that we could come to the conclusion that, e.g., some utilitarian theory provides the correct view of ethics? It is very unlikely that this were the case. One of the main points of Nussbaum's criticism against utilitarian and deontological theories is that their focus of attention is far too limited, and that they consequently disregard

---

[35] Nussbaum 1990, 174.

[36] Nussbaum 1990, 174.

[37] Nussbaum 1990, 182–183. Although the programmes of both Nussbaum and Rawls are "equilibriums", they differ from each other: "Perceptive equilibrium is not the same end as reflective equilibrium; it does not use the same judgments or the same faculties." Nussbaum 1990, 186. For Nussbaum, this refers to the value-ladden nature of all ethical procedures. The final criterion for a decision between different kinds of enterprises is "whether they are capable of doing full justice to everything that our sense of life wants to include." Nussbaum 1990, 186.

[38] Nussbaum 1992, 11.

features essential to a comprehensive understanding of human life. Adopting utilitarianism as a moral theory would, however, necessarily mean restricting one's moral attention to the features essential for that. This would be contrary to Nussbaum's Aristotelian starting point. Hence, utilitarianism and deontological theories do not represent alternative approaches to ethics, in the full meaning of the word, within perceptive equilibrium. They must rather be regarded as points of view which help enrich the picture of human life which is a necessary starting point for ethical study. For Nussbaum, the fault in utilitarianism and deontological theories does not so much lie in what they tell people to do, but in the way they tell people to see their lives. Seen from this perspective, perceptive equilibrium is not as open an ethical method as Nussbaum claims it to be.

What does the preceding analysis reveal of the concept of a person embedded in Nussbaum's theory? As in the theories we have already examined, the moral person of the present theory is defined by the *morally relevant*. Unlike many other theorists Nussbaum deliberately avoids establishing any fixed set of necessary and sufficient conditions here. Consequently, there are no firm criteria for moral personhood either. Everything that is meaningful has importance, and the criteria of meaningfulness cannot be given apart from the particularity of a given situation. Nothing that can occur in the lives of human beings can be excluded from scrutiny. The undefinability of moral personhood has the consequence that all aspects of human life tend to undergo a moral colouring. There is nothing that could *a priori* be left morally unattended, and furthermore, everything that has any meaning has ethical significance as well. Analogical to the open or undefined conception of the moral person is the notion of ethical methodology. In moral thinking all human capacities must be given a role. Ethical deliberation is not just a matter of the intellect but here the emotive and perceptive faculties of human beings also play a significant part. These capacities also characterize the moral person in Nussbaum's theory.

The concept of a person which we have explicated in terms of the morally relevant can be analyzed from a further perspective. It is pre-

cisely this notion of a person which establishes a link between Nussbaum's theoretical starting point and her normative moral theory. The morally relevant as displayed by the concept of a person represents two central aspects of ethics. First, it is an explication of the features of human life to which we have to pay attention in our moral reasoning and of the capacities we have to employ in doing this. The concept of a person is built into the notion of ethical relevance and imported into ethical methodology. Second, the concept of a person serves as a model for good human life. Thus, adopting the perceptive equilibrium not only represents the most appropriate ethical method, but it also offers the best way for living a human life. From this perspective, "moral person" establishes both a methodological device for ethical thinking and defines a normative model for good human life. As such it is not an ethically neutral concept, but a concept that connects the theoretical starting point of a moral theory and the normative theory. If this analysis is a correct one the conclusion we must draw from it is that we cannot separate the theoretical starting point of a moral theory from its normative content. There is always a normative aspect embedded in the theory through the definition of what is morally relevant. Whether this is a conclusion we come to only in relation to Nussbausm's theory is a question we must consider in the last chapter of this study.

3.4. Person as a narrative

Alasdair MacIntyre is known as an outstanding critic of modern Western culture and moral philosophy. The title of his best-known work *After Virtue* refers to his duel aim as moral philosopher; he is both a critic and a reformer. The context in which theoretical ethics, as we know it, is practised is an "after virtue" situation; the ideas of the Enlightenment have replaced the classical Aristotelian-Thomistic scheme of virtue ethics. According to MacIntyre's view, the abandonment of the traditional Western form of moral philosophy has been a great mistake the consequence of which is a deep practical and theoretical crisis. The situation can be corrected, however, if a profound change is effected in ethical thinking. An integral part of the much needed reform is a return to the classical form of ethics. This is the reason for MacIntyre to be writing "after virtue".

In the following, I will explicate MacIntyre's concept of a person by examining three themes in his theory. The first part of the presentation comprises of MacIntyre's criticism of modern morality. The second theme is MacIntyre's version of virtue ethics, especially his narrative conception of ethical thinking. The last issue to be examined is MacIntyre's notion of a tradition-constituted inquiry and the concept of rationality based upon it. The analysis will concentrate on two things. It aims at explicating the theoretical premises of MacIntyre's theory. In relation to this task, the study concentrates on seeking possible connections and analogies between different conceptions in the theory, especially as to how the notions relate to and explicate the concept of personhood. The chapter will close with an examination of some central philosophical themes relevant to the theory.

3.4.1. THE END OF MORAL PHILOSOPHY

In the West today, says MacIntyre, there is nothing that could be called ethics. The field of study known as moral philosophy represents but the last remnants of a vanished culture, senseless ruins of a moral language that has lost its validity in the lives of people. Although moral philosophy as an academic discipline continues and ethical discussion among ordinary people goes on, Western culture has as a matter of fact suffered a moral catastrophe. One sign of this is the lack of almost all means for diagnosing the situation as a disaster.[1] According to MacIntyre's interpretation, there are, however, distinct symptoms indicating that the West actually lives in the midst of a deep crisis. Certain features typical of all ethical discussions and of moral controversies demonstrate this fact. First, it has become very difficult, if not impossible, to argue from an ethical point of view. People represent various rival moral arguments, which are, within themselves, logically valid, but any attempt to relate one rival theory to the other reveals them to be incommensurable with no rational way of weighing their diverging moral claims.[2] Consequently, any moral position appears to be a result of an arbitrary and highly subjective choice unsupported by any criteria beyond the subject. This means that the arguments about value fall outside the scope of rational discourse.[3] As a result, discussion concerning the rationality of the end of an action is reduced to a problem regarding the most efficient means to an arbitrarily chosen end.[4]

The notion of a *moral agent* or a *self* has been exposed to a similar development. Today's moral agent, the *emotivist self*, is an individual who

---

[1] MACINTYRE (1983, 1; 1987, 2–4) claims that moral philosophy has undergone a deep change which has evaporated its basic meaning-giving context. Consequently, our moral language now consists only in fragments of a conceptual scheme for the understanding of which we no longer have the tools. The disquieting feature in this situation is that the methods of philosophical analysis offer no help for recognizing the malady or for finding a panacea. Neither analytical philosophy nor phenomenology and existentialism can solve the problem. This is due to a factor common to all present forms of moral philosophy: they all accept a similar attitude towards their basic conceptual schemes and regard them as universally valid irrespective of time and place. This is a mistake.

is detached from every social setting and communally definable role and who is determined solely by her own likes and dislikes, desires and passions. She deliberates over her course of action by relying on her inner feelings as they happen to occur at the time of decision-making.[5]

Modern ethics is further characterized by its mode of speech. It is commonly accepted that valid arguments in moral discourse are imper-

---

[2] MacIntyre uses the term 'incommensurable' to refer to non-comparability between different theories; not as, e.g., Nussbaum to allude to separate, uncombinable ethical goods or values within a single ethical theory, see page 188. MacIntyre also discusses incommensurability, not just between two or more moral principles within the Western culture, but between whole cultures, or world views. See, e.g., MacIntyre 1990, 4–5. There are all kinds of moral claims which seem to be valid in their own right, but there seems to be no way to weigh them against each other. So, for instance, concepts of utility and justice seem to be mutually incommensurable. This phenomenon is characteristic of Western liberalism: different kinds of evaluation, each independent of the other, are exercised in different types of social environment. The heterogeneity of values prevents any overall ordering of goods. MacIntyre regards the Western societies with pessimism: pluralism which relativizes all value threatens to submerge our culture. MacIntyre 1983, 5; 1987, x, 8, 35, 70, 226; 1988, 1–2, 337.

[3] MacIntyre (1990, 9–12) asks, using the Gifford lectures as his material, whether people who represent different ethical views could agree that progress has been made in the discussion concerning the foundations of ethics during the past century. He comes to the conclusion that there has been no advancement within moral philosophy, but that there are many unresolved disagreements. There is no consensus about the basis of ethics for three reasons. First, there is no agreement on a set of first premises, but a multiplicity and heterogeneity of intellectual traditions rule moral philosophy. Second, there is no agreement on how mutually incompatible considerations should be ranked and no shared standards of value. And finally, there is no common view of rationality, and for this reason, no agreement about where justification of belief ought to begin.

[4] The conflict in demands represented by different interest groups and individuals is not the only characteristic feature of the modern moral debate; the individual self too has become a battlefield of mutually incompatible claims. The roles of the individual form a set of compartmentalized spheres, and each sphere represents some separate good. To solve the conflicts caused by the demands of these different spheres, one then uses the decision procedure which offers the most efficient means for fulfilling one's strongest preferences. MacIntyre 1988, 337. "Choice" is a central concept in modern moral philosophy, and people are constantly affronted with the demand to choose. It is characteristic for the issue that although the demand to choose comprises even people's values, there is no way of justifying one's choice rationally. Evaluation takes place outside the domain of rationality. Consequently, the question about the criteria for rationality concerns only the choice of means. The issue, then, is to select the best, and most effective means for achieving a given end. This approach has made effectiveness modernity's central value. MacIntyre 1983, 9–10; MacIntyre 1987, 25–26; MacIntyre 1988, 337.

sonal, detached from all particularities of the present situation. Accordingly, reasoning is neutral and impartial if it is free from social and cultural particularity. Modern moral philosophy regards this as a necessary condition of non-partiality and of rationality.[6] Furthermore, the situation has become even more complicated because the theoretical difficulties which plague moral philosophy are treated, not as symptoms of a present crisis, but as universal characteristics of every ethical discussion. In philosophy all moral discourse is interminable and all its problems are irresolvable.[7]

Another typical feature of modern morality is that it derives from a plurality of historical origins. This fact has, however, been ignored in modern moral philosophy, and the history of morality is, therefore, treated as if it concerned a single debate with a relatively invariant subject-matter. In reality, moral language has undergone a profound change during the last three hundred years, and during that time the meaning of central moral concepts has changed radically. The real nature of the modern self and the profound difficulties in modern ethical discourse become intelligible only as end-products of this historical process.[8]

A central characteristic in MacIntyre's criticism is his insistence that the faults of modern moral theories are not simply defects which could be corrected by reforming the current models of ethical thinking. On the contrary, a total revision is needed. An integral part of this revision is an exploration of the historical process which has lead to the present situa-

---

[5] MACINTYRE 1987, 8, 19, 30–32; 1988, 213–214. The emotivist self exists only in the present tense: it has no history, no given continuities and no social identity. In accordance with these features, the modern moral agent is seen as capable of criticizing everything and every point of view by virtue of the emotivist self's ability to choose any standpoint for moral criticism and to refrain from adopting any position to form a commitment. MACINTYRE 1987, 31; 1988, 133. For further details concerning the social setting in which the emotivist self is at home in the Western liberalist society, see MACINTYRE 1988, 335–337.

[6] MACINTYRE 1987, 8, 70. Modern ethical discussion has largely adopted its models of explanation from modern natural sciences: human action is explained by facts ordered by some universal generalization. In this view, explaining something equals laying bare the physiological and physical mechanisms which underlie action. No reference is made to purposes, intentions, or reasons for action. MACINTYRE 1987, 82–83.

[7] MACINTYRE 1987, 11.

tion. This historical study is necessary for two reasons. First, it provides a better understanding of the characteristics of the present cultural circumstances. Second, it forms an essential constituent of a sound moral philosophy, because a historical approach is the only way to make sense of ethics.[9] Let us now examine MacIntyre's reconstruction of the history of Western ethics to see what role it has in his theory.

The idea behind MacIntyre's reconstructed history of moral thought is that all the dominant modes of modern moral philosophy are heirs of a common progenitor, the Enlightenment. Despite the differences in these theories, they all aim at a common goal. For them, the task of moral theory is to find a universally justifiable basis for morality, a basis detached from social and cultural particularities.[10] Working from this universally valid metaethical ground, universally acceptable moral rules or principles can then be derived. Although no one has yet succeeded in formulating an ethical theory which fulfils these requirements, the conceptual premises underlying this ideal have not been contested or abandoned, but they still form the common starting point of all prevailing variants of moral philosophy. The perfect moral theory is universally valid, applicable to all situations irrespective of time and place. The philosophical conclusion which has been drawn from the failure of the Enlightenment project of moral philosophy has been that evaluative positions or convictions cannot be supported by rational arguments. All current forms of moral theory accept this view without reservation.[11]

---

[8] "A key part of my thesis has been that modern moral utterance and practice can only be understood as a series of fragmented survivals from an older past and that the insoluble problems which they have generated for modern moral theorists will remain insoluble until this is well understood." MacIntyre 1987, 110–111. See also MacIntyre 1987, 10–11, 35. The development within moral philosophy is not just any kind of change; MacIntyre suggests that the motivation behind it is manifold , see, e.g., MacIntyre 1987, 14–18, where he examines the background presuppositions of the emotivist culture.

[9] MacIntyre 1987, 2. MacIntyre maintains that analytic philosophy and its model of explanation has effectively ingored the historical nature of the problems in moral philosophy. Analytic philosophy has claimed to represent a universal and timeless form of practical reasoning, where it has, in fact, been a form of reasoning specific to its own social and political background, which is the liberal individual culture. MacIntyre 1988, 340.

[10] MacIntyre 1988, 6.

According to MacIntyre's interpretation, the failure of this philosophical enterprise derives from a misconception concerning the nature of morality. The only way to make sense of any moral system is to see it against the background of the particular human community, its history and practices, in which it has evolved. It is a mistake to seek a universally valid basis for ethics and to try to formulate universal moral principles; morality is a historically determined human institution, and can be understood only if it is seen as such.[12] The Enlightenment project faltered because there is simply no way of understanding human life and action, and hence morality either, outside a history, a tradition, and a context all of which provide it with the criteria of meaning and rationality.[13] The phenomena of human life do not make sense without a framework of belief to lend meaning to particular actions. Ethics cannot be equated with some abstract moral rules detached from the specific situation of people who live in particular communities.[14]

To analyze MacIntyre's criticism in more detail, we can say that his approach challenges one of the key premises of modern ethical thought, i.e., the idea of *universalizability*. He abandons the conception of a univer-

[11] "[...] the legacy of the Enlightenment has been the provision of an ideal of rational justification which it has proved impossible to attain. And hence in key part derives the inability within our culture to unite conviction and rational justification. Within that kind of academic philosophy which is the heir to the philosophies of the Enlightenment enquiry into the nature of rational justification has continued with ever-increasing refinement and undiminishing disagreement." MACINTYRE 1988, 6. MacIntyre maintains that Nietzsche's conclusion in his *Zur Genealogie der Moral*, according to which all rational vindications of morality manifestly fail, was correct, although not of morality as a whole, but as a statement concerning the post-Enlightenment moral philosophy. MACINTYRE 1987, 117. One of the tasks MacIntyre has set himself is to show that Nietzsche's thesis does not cover all ethical thinking, that its criticism only hits the moral philosophy of the last three hundred years, that the projects of Aristotle and Thomas Aquinas can face the challenge of Nietzsche and his existentialist and emotivist followers. MACINTYRE 1987, 117–118. See also a longer discussion about the controversy between these rival theories of moral enquiry in MACINTYRE 1990.

[12] MACINTYRE 1987, 67, 126–127, 146.

[13] MACINTYRE 1987, 51–52; MACINTYRE 1988, 6–7. FRANKENA (1983, 580, 582) points out that MacIntyre cannot arrive at his conclusion concerning the failure of the Enlightenment project by means of historical study, but that his arguments are drawn from the analytical philosophy which he so strongly criticizes.

[14] MACINTYRE 1987, 59, 118, 126.

sally valid starting point for moral thinking and, consequently, the idea of universally valid moral precepts. The point of MacIntyre's criticism is not, however, that modern moral theories tell people to do wrong things, but that they represent a mistaken view of the nature of ethics. Furthermore, these theories employ, in MacIntyre's view, a misconception concerning the notions of rationality and of human life. What then would be the correct understanding of these concepts? For MacIntyre it is a *universal feature* of human life, morality and rationality that they are *particular* and *historically determined*. What is noteworthy in MacIntyre's criticism of the heirs of the Enlightenment is that he fails to notice how the idea he is criticizing in others comes to occupy a central role in his own argument. MacIntyre does not abandon the notion of a universal basis of ethics in his own thinking, he only adopts a new way of determining it. For him, the historical and particular nature of human institutions is the universally valid starting point upon which all sound understanding of ethics bases.

In this connection, we can also ask whether MacIntyre's theoretical starting point can serve as a basis for the concept of a person in his theory. If it can be so conceived then moral personhood must be understood against the framework of a person's particular, historically determined community with its mores and practices. We cannot, however, determine whether this interpretation is correct at this point of the study, but must return to the question later.[15]

Despite the mutual search for a universal theoretical starting point both in MacIntyre's theory and in modern moral theories, there is one respect in which MacIntyre's model deviates from those theories. In MacIntyre's version of ethics there are no universally valid normative rules. If life is universally particular it is not likely that there exists universal normative principles. This has the effect that it is no longer meaningful to concentrate on examining rules in ethics. Instead attention should be paid to human life from a wider perspective.

MacIntyre's view of the nature of ethics and the universal in morality

---

[15] See page 216.

is a theoretical position which primarily concerns the manner of practicing moral philosophy. But does this theoretical position have normative implications? MacIntyre's conception of the task of moral philosophy is similar to that of the post-Enlightenment moral theories. All these theories offer a theoretical framework for understanding the institution of morality. All theorists we have examined in the course of this study, MacIntyre among them, stress the ethical importance of their theories. They seem to think that a theoretical approach to morality, which is to say the correct view each of them claims to represent, has an impact on normative ethics although they do not explicate *what* the nature of this connection is. Here MacIntyre is no exception; he is very emphatic that only a correct view of the nature of ethics can produce a sound morality. MacIntyre's theoretical starting point is not a *moral* position; and yet neither is it ethically "neutral" for it determines the perspective for the moral, and can thus have ethical significance.[16]

Let us return to MacIntyre's restoration programme. MacIntyre has set himself a twofold task. First, he constructs the history of moral thought to highlight the state and problems of modern moral philosophy. Using this analysis as a basis, he then introduces an alternative mode for doing moral philosophy. These two aspects of the programme go together, for MacIntyre employs the same conceptual framework both for laying bare the history of moral philosophy and for modelling an alternative moral theory. Consequently, one cannot agree with the historical reconstruction unless one also accepts the conceptual model for understanding ethics. In other words, MacIntyre's historical mode of approach and his ethical restoration programme go together.[17]

---

[16] VON WRIGHT's (1968, 6) short remark seems relevant in this connection: conceptual investigation of moral terms, usually classified as metaethical, is actually normative in the sense that it sets out standards which we use for our orientation in the world as moral agents, and which aim at directing our lives.

3.4.2. Ethics of virtue and a narrative concept of a person

How should we, then, correct the mistakes of the Enlightenment project and improve the condition of moral philosophy? The failure of the Enlightenment endeavour and the false axioms which led to it have eroded the connection between rationality and morality and the link between facts and values. MacIntyre's main goal is to re-establish this link. Morality cannot make any sense as long as the sentence: "it is rational to be good" has no validity in our conceptual system.[18] In other words, MacIntyre aims at developing a new conception of practical rationality which would make rationality and morality coincide. But this is, as a matter of fact, the explicit aim of all the "post-Enlightenment" moral theories we have examined in this study, whether utilitarian or contractarian. From this perspective, MacIntyre's theory does not establish anything radically different. His conceptual framework may be different, but his theoretical aim is the same as that of other moral theorists: to define a meaningful notion of practical rationality.

MacIntyre develops his conception of practical rationality in three stages which eventually lead to a reinstated ethics of virtue.[19] The first task is to construct an understanding of ethics which links together the notion of a human being and the meaning of moral concepts. The second requirement is a *teleologically* conceived morality. This involves an idea of change and development embedded in the concept of a good human life. Here ethics is understood as the bond between human nature as it is

---

[17] MacIntyre points out that controversies like the one between modern moral philosophy and his own theory are particularly exacting, while disputes between differing conceptual frameworks include disagreements about the conception of rationality. There is no neutral ground for evaluation and comparison between the two, but the very criteria of rationality upon which each party of the controversy attempts to ground its own arguments and reasoning are part of the dispute. So, it is characteristic of the Enlightenment-type of rationality to seek timeless and impartial criteria for rationality, whereas such a search is, from MacIntyre's point of view, rational only within a very limited social context; the criteria for rationality, and even the concept of rationality have to be regarded as relative to some historical point of view. MacIntyre 1988, 4, 7, 9.

[18] MacIntyre 1987, 186.

[19] MacIntyre 1987, 186–187.

and as it should be so as to be good. The third step in MacIntyre's project is to show how different and mutually incompatible, and therefore rival systems of morality can be rationally compared with and weighed against each other.

The study now presents the three elements upon which MacIntyre builds his ethical theory in more detail. First, it examines the functional aspect of the theory and then pays attention to the teleological character of the project. In this connection, the analysis takes up some questions central to MacIntyre's substantive moral theory. The last part of the present section deals with the question as to which moral system we ought rationally to adopt.

Is the distinction between facts and values absolute? Could we not find any uses of evaluative terms which connect facts with values? MacIntyre maintains that there are such uses. When we ask what, say, a good farmer is like, the set of criteria we give as an answer is factual, composed of a *functional description* of what it involves to farm well. Actually, we cannot even define what farming is without including the criteria of good farming in the definition. Now, if we regard expressions indicating moral evaluation from a similar perspective, as parts of descriptions in functional terms, the gap between facts and values disappears.[20]

Bridging the gap between facts and values involves a return to the Aristotelian and the mediaeval ethical tradition in which *man* was a central functional concept. In this tradition, man is not conceived of as an individual disconnected self, but as an agent rooted in the forms of social life, in its many socially and culturally determined roles and tasks,

---

[20] MACINTYRE 1987, 57–58. "Within the Aristotelian tradition to call x good (where x may be among other things a person or an animal or a policy or a state of affairs) is to say that it is the kind of *x* which someone would choose who wanted an *x* for the purpose for which *x*'s are characteristically wanted. [...] The presupposition of this use of 'good' is that every type of item which it is appropriate to call good or bad — including persons and actions — has, as a matter of fact, some given specific purpose or function. To call something good therefore is also to make a factual statement. [...] Within this tradition moral and evaluative statements can be called true or false in precisely the way in which all other factual statements can be so called." MACINTYRE 1987, 59. MacIntyre's solution to building a link between factual and evaluative statements resembles that of Wallace's; see page 181.

through which the life of a human being has a functionally characterized point and purpose. In this context, it is a factual statement to call someone a good person, for a good person is one who fulfils her roles and tasks well in a community sustained by social practices and networks formed by people acting in different roles.[21]

In MacIntyre's theory the context of social life and its different forms also serve another purpose; they constitute an interpretative background to emotions and desires. In the post-Enlightenment moral philosophy, MacIntyre maintains, emotions and desires are regarded as psychic phenomena disconnected from the outer reality. Hence, desires and emotions are conceived as psychologically basic items, largely invariant in their function between cultures.[22] This view is false. Emotions and desires are always linked to the social life of a community, and to the specific roles and forms of action typical of it: emotions and desires are norm-governed. Ignorance concerning these norms necessarily causes an inability to understand the context of interaction and interpretation within which the emotions and desires make sense.[23]

The second point in MacIntyre's moral programme is a conception of morality which is teleological; the functionalistic use of moral terms pre-

---

[21] "[...] moral arguments within the classical, Aristotelian tradition — whether in its Greek or its medieval versions — involve at least one central functional concept, the concept of *man* understood as having an essential nature and an essential purpose or function; and it is when and only when the classical tradition in its integrity has been substantially rejected that moral arguments change their character so that they fall within the scope of some version of the 'No "ought" conclusion from "is" premises' principle. That is to say, 'man' stands to 'good man' as 'watch' stands to 'good watch' or 'farmer' to 'good farmer', within the classical tradition. [...] [The use of 'man' as a functional concept] is rooted in the forms of social life to which the theorists of the classical tradition give expression. For according to that tradition to be a man is to fill a set of roles each of which has its own point and purpose: member of a family, citizen, soldier, philosopher, servant of God. It is only when man is thought of as an individual prior to and apart from all roles that 'man' ceases to be a functional concept." MacIntyre 1987, 58–59.

[22] MacIntyre 1988, 21. The post-Enlightenment way of conceptualizing and understanding passions as part of nature defined independently of culture rather than as an expression of culture, is itself a part of the evaluative system that establishes the meaning-giving framework for emotions and desires in the social world structured by the post-Enlightenment values and ways of thinking. MacIntyre 1988, 77.

[23] MacIntyre 1988, 76.

supposes a teleology. A functionalistic definition of human personhood implies a description of the constituents of a good human life. This definition, interwoven into the fabric of beliefs which give meaning to the practices and customs of a community, also expresses the criteria for what it is to actualize a good human life within that particular community. To be a member of the human species is not equivalent to being a human person in the full sense: to be a human person presupposes the acceptance of socially determined criteria for good human personhood. People are not good human beings naturally, but they need tutoring, education, and advice to develop as persons. In this scheme, morality is given the task of a tutor: it makes it possible to build a bridge between the human state as it happens to be and the human *telos*. Adopting a teleological form of morality necessarily involves a return to the ethics of virtue. The process of transition comprises a change in the person, and to acquire virtues is, on the one hand, a means for achieving the human telos, on the other hand, an end in itself.[24] Interestingly, MacIntyre gives morality a task which is *universal* and independent of any particular situation or historical setting. Morality, in the form of virtue ethics, must be conceived as a tutor moulding "natural" human beings into moral persons. Here again MacIntyre fails to notice that his view of the nature of morality implies a commitment to a certain theoretical position which he regards as universal and independent of historical change.

How does ethics then enable a human being to achieve her communally defined telos? Which form does a morality compatible with the functionalist and teleological model of ethics take? Two issues are central in answering these questions: first, the teleological scheme for understanding human life presupposes that we conceive of it as a *narrative unity*, or, that we think of it as a story.[25] Second, to preserve the functional nature of ethical concepts, a life must be regarded as a whole structured by communally determined *practices*. Practices, for their part, cannot be exercized without virtues which enable people to achieve goods internal to the respective practices.[26] In the following, I will first examine the

[24] MacIntyre 1987, 52–55. 187, 218–219.

concept of a narrative order of life and then concentrate on MacIntyre's model of socially constitutive practices sustained by virtues.

It is worth noticing here that the concept of a narrative unity in life is analogous to the form of MacIntyre's ethical project, i.e., to his historical approach to morality. As we have seen, MacIntyre maintains that ethical concepts do not make sense if they are detached from the historical and social setting in which they have emerged. Now, the same applies to a human person's life: a person must see herself, not as an individual self, but as a participant in a story that neither starts nor ends with her individual birth or death.[27] Again, for the purpose of understanding one's life, it is necessary that one sees one's present self as constituted by one's past history.[28]

The historical development and the particularity of an ethical tradition give us the necessary background for understanding morality in a conceptually correct way. An analogical claim is true of an individual life:

---

[25] MacIntyre 1987, 216. Two things are typical of a human life conceived and made intelligible by the narrative form: the life, which is to say the story of that particular life, is unpredictable, and teleological in form. At no point in a life can we certainly and precisely say what will happen next. The unpredictability is, again, linked together with a certain teleological character which MacIntyre illustrates as follows: "We live out our lives, both individually and in our relationships with each other, in the light of certain conceptions of a possible shared future, a future in which certain possibilities beckon us forward and others repel us, some seem already foreclosed and others perhaps inevitable. There is no present which is not informed by some image of some future and an image of the future which always presents itself in the form of a *telos* or of a variety of ends or goals — towards which we are either moving or failing to move in the present. Unpredictability and teleology therefore coexist as part of our lives; like characters in a fictional narrative we do not know what will happen next, but nonetheless our lives have a certain form which projects itself towards our future. Thus the narratives which we live out have both an unpredictable and a partially teleological character." MacIntyre 1987, 215–216. See also MacIntyre 1987, 124, 187.

[26] MacIntyre 1987, 190–193.

[27] "[...] the story of my life is always embedded in the story of those communities from which I derive my identity. I am born with a past; and to try to cut myself off from that past, in the individualist mode, is to deform my present relationships. The possession of a historical identity and the possession of a social identity coincide. [...] What I am, therefore, is in key part what I inherit, a specific past that is present to some degree in my present. I find myself part of a history and that is generally to say, whether I like it or not, whether I recognize it or not, one of the bearers of a tradition." MacIntyre 1987, 221. See also MacIntyre 1987, 146.

the narrative unity of a life, or a life conceived as a story, grants us the conceptual tools for making sense of human life and action.[29] The way in which human life is defined here shows that the suggestion made earlier[30] was correct. As we have seen, the particular and historically specific conditions in which human beings live determine their ethical thinking. Now we also notice that MacIntyre actually defines the moral person by using the same description: the particular historical context defines human personhood. The central theoretical constituents of MacIntyre's ethical theory are analogous to the concept of a person as explicated from his normative notion of a person.

The framework of the particular also plays a central role in MacIntyre's notion of human behaviour. The conception of behaviour becomes intelligible only through the concept of intention. Intentions,

---

[28] MacIntyre criticizes the post-Enlightenment conception of personal identity. All attempts to give an account of personal identity solely in terms of psychological states or events will fail. The same is true of theories which try to establish a connection between those states and events and strict identity understood in terms of *Leibniz's Law.* The failure is due to the lack of a background against which the concept of personal identity becomes intelligible. The necessary background is provided by the concept of a story and the kind of unity of character which a story requires. To clarify his position, MacIntyre refers to an example: Someone attempts to commit or commits suicide. To explain her deed, the person tells that her life is meaningless, that the story of her life has become unintelligible to her, that it lacks any point, or movement towards a *telos.* MACINTYRE 1987, 217. Here it seems, however, that MacIntyre actually treats what he calls *Leibniz's Law* as a positive principle of the form: "If two things are identical, then whatever is true of the one is true of the other." As a matter of fact, Leibniz only formulated his principle of the identity of indiscernibles in a negative form, i.e.: "It is not true that two substances may be exactly alike and differ only numerically, *solo numero.*" *Discourse on Metaphysics* (IX). In its negative formulation the principle does not lay down the conditions which any two things, *A* and *B*, must satisfy for them to be identical, but describes the conditions which must be satisfied when any two things are *not* identical. See STROLL 1972, 122.

[29] MacIntyre sees two obstacles which inhibit a view of a human life as a whole. The first obstacle is social: human life is segmented into a variety of periods of age, according to different kinds of activities. Nothing binds these different elements into a unity. Second, despite their other differences, both analytical philosophy and existentialism disintegrate human life. Analytical philosophy tends to think atomistically about human action and to analyze complex actions and transactions in terms of simple components; beside these there is no overall point of view. Existentialism, for its part, tends to separate a human being from her roles which makes it impossible to conceive a person's life as a unity. MACINTYRE 1987, 204–205.

[30] See page 210.

for their part, become intelligible against the background of historically and socially determined settings which make those intentions intelligible both to agents themselves and to others.[31] Without the conception of an intention and outside a meaningful framework of belief it is impossible to explain human behaviour.[32] This means that people identify a particular action as a part of two larger contexts. Action is, first, seen as stemming from the intentions of its agent, in a causal and temporal order which derive from the role the particular action has in the person's history. The second essential part for understanding human behaviour consists in the role actions have in the larger history of the communal setting or settings to which they belong. In this way, narrative history is the basic and essential genre for the characterization of human actions.[33]

A human life and the life of a community, could not form a narrative if it were composed of single actions. On the contrary, human life often consists of and is structured by highly developed social *practices* and institutions.[34] A practice is

---

[31] MacIntyre 1987, 206–207.

[32] MacIntyre 1987, 208.

[33] MacIntyre 1987, 208. MacIntyre strongly criticizes the modern models for explaining human action: human action is quite unintelligible when it is understood in terms of Humean and post-Humean accounts in which truth-values can only be attached to statements of fact. What is needed is the Aristotelian model for practical reasoning. Aristotle's account has four essential elements. First, the wants and goals of the agent provide the context for the reasoning. The second element needed is a belief that having, doing, or seeking such-and-such is the type of thing that is good for or needed by a so-and-so; this assertion also forms the major premise in the agent's reasoning. The minor premise, i.e. the third element required in practical reasoning, is the agent's perceptually acquired assertion that the instance at hand is of the requisite kind. The fourth element in practical reasoning, and this is the conclusion of the practical syllogism, is the action towards the desired end, for the attainment of some good. MacIntyre 1987, 161–162.

[34] MacIntyre 1987, 187. Practices and institutions are not to be equated with each other, although they are closely related. Institutions are the bearers and sustainers of practices: they ensure, if they function well, the external goods necessary for the realization of the respective practices; hence, for instance the institution of hospitals is to secure the practice of good medicine. Institutions may also operate against a practice: in such cases the maintenance of the external goods with the support of which the practice would flourish, will become an end in itself. MacIntyre 1987, 194.

> any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended.[35]

Two features characterize practices. To participate in a practice involves achieving goods internal to that practice; i.e. those who take part in a practice get a "share" of a good which does not exist outside the particular practice, but which the practice and participation in it constitute. Second, these goods can only be identified and recognized with experience gained from participation in the specific practice. Thus, inexperience in regard to a certain practice can be equated with incompetence to judge the quality of that practice.[36]

All practices have a history. They have undergone a process of development and change; and this development is perpetual for any vital practice. Accordingly, there are certain historically determined *standards of excellence* and a specific set of rules which lay out and structure the achievement of goods. To enter a practice, or to become a partaker, involves acceding to a subordination to the historically formed standards of the practice and accepting one's own noviceship.[37] Practices require the exercise of certain technical skills characteristic of each of them; but besides the actualizing of certain specific talents they necessarily require *virtues* of their practitioners. A virtue is an acquired human quality the possession and exercise of which tends to enable people to achieve goods internal to practices and the lack of which incapacitates people

---

[35] MACINTYRE 1987, 187. It is not simple to draw a line between activities which form, and others which do not form a practice. MacIntyre gives some examples which clarify the distinction. Thus, throwing a football with skill is not a practice, but the game of football is; bricklaying is not a practice, but architecture is; planting turnips is not, but farming is a practice; and so forth. MACINTYRE 1987, 187–188.

[36] MACINTYRE 1987, 188–189; MACINTYRE 1988, 122–123.

[37] MACINTYRE 1987, 190.

from achieving such goods.[38] The connection between virtues and the internal good of practices is so close that people who do not cultivate the virtues inherent to the practices will never acquire the goods internal to practices. This means that the cultivation of virtues constitutes the goods internal to practices.[39]

But is it possible to build a system of ethics solely on virtues and goods tied to practices; is there no place for rules and moral norms in MacIntyre's ethics? Virtues alone are not enough to secure the good of a community and the good of a human being: rules and norms are a necessary part of a moral theory. But the function of rules has to be different from the role that rules have acquired in post-Enlightenment philosophy: rules are not to be detached from the good of a moral agent. Moral rules can be integrated to the good of the agent only within an Aristotelian type of virtue ethics. This form of moral theory prohibits certain acts absolutely, irrespective of circumstances and consequences. The nature of these acts is such that a virtuous person would never commit them, because becoming guilty of the offences they prohibit would destroy the relationships which make common pursuit of the good possible in one's community. If the necessary conditions for realizing the

---

[38] MacIntyre 1987, 191. MacIntyre 1987, 219. See also Murdoch 1967, 17–18, 20 for a similar view.

[39] MacIntyre 1987, 193. In the postscript of the second edition of *After Virtue* MacIntyre clarifies some of his central points in answer to his critics; he writes about the relation between goods and virtues: "The importance therefore for beginnning [sic] from practices in any consideration of the virtues is that the exercise of the virtues is not only worthwhile for its own sake — it turns out that you cannot be genuinely courageous or just or whatever without caring for those virtues for their own sake — but has further point and purpose, and indeed that it is in grasping that point and purpose that we characteristically intitially [sic] came to value the virtues. Yet the virtues are not related to the goods which provide them with further point and purpose in the way in which a skill is related to the ends that its successful exercise procures or in the way in which a skill is related to those objects of our desire that its successful exercise may enable us to possess. [...] The goods internal to practices which cannot be achieved without the exercise of the virtues are not the ends pursued by particular individuals on particular occasions, but the excellences specific to those types of practices which one achieves or fails to achieve, moves toward or fails to move toward in virtue of the way in which one pursues one's particular ends or goals on particular occasions, excellences our conception of which changes over time as our goals are transformed." MacIntyre 1987, 273–274.

common good were weakened, the common life of the community would be endangered. Again, the basis of life in a community will be destroyed if there is no common good to pursue. Consequently,an ethics of virtue must be supplemented by some account of those types of action which are absolutely prohibited. [40]

MacIntyre does not deal with the question of moral rules in detail. Hence, it is not possible to determine whether he thinks that there could be universally valid moral rules. From what he says we can, nevertheless, infer a universally valid *formal* criterion for moral rules. All actions which destroy social relations necessary for the pursuit and maintanance of the historically determined good of the community are absolutely prohibited. As such, the formulation resembles the conditions social contract theories set for the ethically acceptable.

Let us return to virtues and regard them from another, wider angle. Seen in the context of the narrative unity of a life the meaning of virtues as part of a person's whole life becomes intelligible. From this more extensive point of view, virtues direct a person's behaviour towards attaining the goods internal to human life.[41]As was noticed above,[42] the narrative form of a life presupposes a *telos* towards which life is directed. The telos of a human life is the search for a good life for a human being. In this a good human life means "a life spent in seeking for the good life for man".[43] Now, the virtues occupy the same role in relation to the unity of a life as they have in regard to practices: their task is to ensure the achievement of goods which are constituted by the exercise of virtues. To sum up, a good human life is simply a search for the good. The form

[40] MACINTYRE 1987, 150–152.

[41] MACINTYRE 1987, 201–203, 219–220.

[42] See page 214.

[43] "[...] because my life is to be understood as a teleologically ordered unity, a whole the nature of which and the good of which I have to learn how to discover, my life has the continuity and unity of a quest, a quest whose object is to discover that truth about my life as a whole which is an indispensable part of the good of that life. So on this view my life has the unity of a story with a beginning, a middle, and an end, beginning with birth and ending, so far as concerns the final judgment to be passed on it — in respect of the achievement of my good — with death." MACINTYRE 1990, 197. See also MACINTYRE 1987, 219.

of this search is a story, and it is structured by practices within which virtues facilitate the partakers to acquire the goods internal to these practices. This means that virtues also have a function which is independent of practices; they enable a person to discern what is good for her *telos* as a human being.[44]

We have examined the two first points in MacIntyre's ethical programme. It is now time to move to the third part and to analyze the possibility of comparing different, mutually incompatible moral systems with each other. Such comparison is necessary for MacIntyre's concept of practical rationality, for his theory claims there are no universal concepts but that even our conceptions of rationality are historically particular. Hence, to avoid sinking into total relativism, there has to be a way for comparing and evaluating rival moral systems.

[44] MacIntyre 1987, 219–220, 273.

3.4.3. Social narratives — tradition as the context of the ethical

Good human life takes place in a *social tradition*. The story of a person is always embedded in the narrative of the communities from which her identity derives. This socially structured and constituted identity is not an object of choice, for all people are born with a past. A person becomes annexed to this intellectual, cultural and philosophical past as well as to the moral history of a particular social tradition at birth. Thus, inherited particularity is an inevitable condition of human personhood. This does not, however, imply that anyone is "condemned" to accept the moral and other limitations of this specific particularity: it constitutes only the necessary starting point of the quest for a good life.[45]

MacIntyre defines a tradition as an argument extended through time: it includes a specific mode of reasoning, it relies on particular criteria of rationality, and it requires a set of virtues distinctive of it. Traditions actualize in history in communities which share what could be called a common *form of life*. The practices and institutions of a social community display the distinctive beliefs, values, etc., embedded in its specific tradition. The only background against which practices and institutions can have a meaning, and in relation to which they make sense is the good the pursuit of which gives its particular point and purpose to that tradition.[46]

The relation between a tradition and the virtues specific to the practices which lend a certain social form to that tradition are twofold. On the one hand, a tradition cannot survive if people who live within it do not exercise virtues which sustain it. On the other hand, these virtues cease to make sense if the meaning-giving horizon of a tradition vanishes: they simply become forms of behaviour detached from all evaluative criteria.[47]

---

[45] MacIntyre 1987, 220–221.

[46] MacIntyre 1987, 222–223; 1988, 12–13. MacIntyre does not use the term "form of life" but I apply it here to refer to the social and cultural expressions of a tradition in the life of a community.

[47] MacIntyre 1987, 222–223.

Interestingly, the relation between a tradition and its specific practices is analogous to the relation between a moral tradition and the concept of good, which has been examined earlier.[48] As was the case with the concept of good, moral tradition here establishes a framework which locates particular practices within a functionalistic scheme. This, again, makes it possible to re-establish the link between morality and rationality so that a connection can be found between values and facts: moral virtues become intelligible against the background of practices which sustain the particular community. In addition, the concept of a tradition functions as a justificatory ground for ethics because moral concepts form an integral part of the tradition in terms of which the members of the community understand their lives. Hence, a tradition establishes a meaning-giving framework in two senses. First, it is a system of thought within which moral language becomes intelligible. Second, it is a conceptual structure enabling the members of a tradition to find meaning and purpose in their lives. This twofold role of the notion becomes apparent in the context of practising virtues: in order to achieve the goods inherent to practices one has to practise certain virtues. Apart from the goal of a particular practice, this activity also sustains the tradition, and as such it enables the members of the community to participate in the goods inherent, not only to the particular practice, but to their tradition.

In the present context it is useful to examine yet another theme relevant for MacIntyre's concept of a social tradition.[49] According to him, it is characteristic of all vital and living traditions that they embody *continuities of conflict*. Even the widely accepted modes of thought, criteria of rationality, and the current moral norms accepted within a tradition, are defined and redefined in terms of conflict. All agreements and central conceptions of a tradition are an outcome of conflict, or of a series of debates in terms of which the central epistemological and other commitments of the tradition have been advanced, evaluated, modified and

---

[48] See page 209.

[49] The following discussion does not directly concern the concept of a person but is. nonetheless necessary to the present study if we want to understand MacIntyre's theory in its totality.

accepted. Other views have, again, been rejected in similar debates.[50] MacIntyre distinguishes three stages in such controversies, which together form the *tradition-constituted* and *tradition-constitutive enquiry*.

This kind of an enquiry begins in and from some condition of pure historical contingence. Hence, the beliefs, institutions and practices of some particular community constitute the given of the inquiry. In the first stage of the process, the community's sources of authority receive an unquestionable and firm status. Such authority can be ascribed to certain beliefs, utterances, sacred texts or documents on the basis of their exceptionality and meaningfulness for the community. The authority can also become manifest in persons who occupy central roles and positions in the community.[51] This stage, which can be characterized as a peaceful *status quo*, may be only imaginary, even in MacIntyre's own terms. A living tradition is, MacIntyre often stresses, always in a state of change. It would, then, be better to say that this description is more applicable to some part of a tradition than to a tradition as a whole.

In the second stage of the process the central premises, upon which the beliefs, practices and institutions of the community lie, are shown to be somehow susceptible. This may derive from an explication of hidden incoherences which are situated in the established system of belief; or it can turn out that the tradition is lacking in resources for solving an urgent problem. Irrespective of the origin of such conflict, the situation always gives rise to an *epistemological crisis* in the tradition.[52] The tradition and its ruling agreements must then face a confrontation with new situations and new questions. Sometimes the tradition is challenged by critics and enemies who stand outside it and who reject all, or at least key parts of its fundamental premises. But there are also conflicts which stem from within the tradition. These are debates between mutually incompatible forms of argument arising from the history of the tradition itself.[53]

---

[50] MACINTYRE 1989, 350, 354.

[51] MACINTYRE 1989, 354.

[52] MACINTYRE 1988, 355, 361.

[53] MACINTYRE 1988, 12; 1988, 354–355.

The process moves into the third phase when an attempt is made to solve the inadequacies detected in the tradition. It may become apparent that the tradition is lacking in resources so that it cannot overcome the summons it faces. This means that the tradition ceases to make progress by its own standards of progress, and it fails in the task which the tradition itself has set.[54] Solving an epistemological crisis becomes possible only by enriching the tradition with new concepts, which involves imaginative conceptual innovation and formulation of a new type or types of theory. Besides this innovative enlargement of the conceptual framework the solution must also include an explanation of the reasons for rendering the tradition insufficient before the challenge which lead to the crisis. Furthermore, the solution has to exhibit some fundamental continuity with the conceptual and theoretical structures of the tradition before and after the epistemological crisis.[55]

Such disputes can arise, not just between differing moral arguments but also between distinct social traditions. Moral conflicts can, in fact, be just symptoms of more profound divergences, such as separate systems of belief or differing accounts of rationality. Thus, for example, the issue between two mutually incompatible moral claims can lie, not in the two different moral positions, but in two mutually irreconcilable traditions. Here the real question is not how to settle the dispute between the two moral claims but how to compare and contrast the two traditions with each other; and this can only be done through a process of historical enquiry.[56]

---

[54] MacIntyre 1988, 355, 361–362.

[55] MacIntyre 1988, 362–365. MacIntyre's description of the three stages of an epistemological crisis of a tradition is analogous with Thomas Kuhn's famous theory of scientific revolutions, Thomas Kuhn, *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago, 1962. A tradition in MacIntyre's theory corresponds to a scientific theory in that of Kuhn's. The peaceful *status quo* of a tradition is analogous to the concept of normal science. The description of disruptive threats to the ruling paradigm of a tradition/science is similar in both theories. On the third phase of the process, described as overcoming an epistemological crisis and formation of a new scientific paradigm, however, MacIntyre's point of view is somewhat different from that of Kuhn's. Although conceptual innovation is central in both models, MacIntyre focuses on continuities between the earlier and the later tradition, whereas Kuhn interprets the development as a scientific revolution in which the new theory completely replaces the old.

All people live within some social and moral tradition. They have grown into their own tradition, into its practices and institutions, into its systems of belief, and they have adopted its account of rationality. Their particular moral tradition is something they have accepted as given and not acquired as a result of an individual choice. Besides, as partakers of a tradition they have no means of adopting general, timeless standards with the help of which they could ascend above the particularity of their situation or above that of others.[57] If this is true of everyone's situation how is it possible, MacIntyre asks, to evaluate and justifiably criticize a foreign tradition, the forms of life and criteria of rationality of which we do not share? Does not the view necessarily lead to relativism? Further, if there are no universal criteria for evaluing moral traditions, and if even all criteria of rationality are relational to traditions in general, and to a certain moral tradition in particular,[58] does it not become obvious that participants from different traditions can neither formulate nor discuss points in which their traditions differ from each other? Must we not admit that one moral view is not better than another, but that all different points of view simply display varying but equally valid perspectives on moral questions? These two types of counterargument form the *relativist challenge* and the *perspectivist challenge* to MacIntyre's view.[59] MacIntyre maintains that his theory copes with both types of difficulties and that his tradition-bound ethical theory enables both to combine morality and

---

[56] MACINTYRE 1987, 269.

[57] "The conclusion [...] is not only that it is out of the debates, conflicts, and enquiry of socially embodied, historically contingent traditions that contentions regarding practical rationality and justice are advanced, modified, abandoned, or replaced, but that there is no other way to engage in the formulation, elaboration, rational justification, and criticism of accounts of practical rationality and justice except from within some one particular tradition in conversation, cooperation, and conflict with those who inhabit the same tradition. There is no standing ground, no place for enquiry, no way to engage in the practices of advancing, evaluating, accepting, and rejecting reasoned argument apart from that which is provided by some particular tradition or other." MACINTYRE 1988, 350. See also MACINTYRE 1987, 268; MACINTYRE 1988, 401; MACINTYRE 1990, 5–6.

[58] When discussing the criteria and canons of rationality within a certain tradition MacIntyre applies the terms *tradition of enquiry*, or *tradition of moral enquiry*; see, MACINTYRE 1988, 363. I will not make similar specifications here. Instead I will only use the term "tradition" and include all the different aspects within the term.

rationality and to reinstate ethical realism.

The tradition-constituted method of enquiry is a debate which is historical and dialectical in form. This discourse can take place only if two difficulties can be overcome: first, the basic commitments of a tradition must somehow *be translatable* from the language of one tradition into the language inherent to another, differing tradition. Second, the discourse is not possible without a concept of *historically determined criteria for rationality.* Unless such a concept is found, true evaluation between dissenting traditions is inconceivable. Of these two the concept of translatability is more important. If there is a way to translate the basic notions and commitments of one tradition into the language of another, the question concerning differing conceptions of rationality can be settled.

What does understanding another tradition mean? Different traditions can be translucent to each other, at least partly, but this circumstance does not necessarily prevail.[60] Sometimes two traditions are mutually incommensurable, in which case understanding the other presupposes translation. MacIntyre distinguishes two distinct species of translation, namely, translation by *same-saying* and translation by *linguistic innovation*, which is more important for the present discussion.[61] Linguistic innovation is necessary in a translation the aim of which is to understand an alien, rival tradition. The reason for this is that the things described in the language of the tradition *A* do not exist in the language of the tradition *B* into which the translation must be made. A member of the tradition *B* can accomplish this only by learning the tradition *A* and its language. The language of the tradition *A* is a second language for the person at home in the tradition *B*, but now this language has to be

---

[59] "The relativist challenge rests upon a denial that a rational debate between and rational choice among rival traditions is possible; the perspectivist challenge puts in question the possibility of making truth claims from within any one tradition. For if there is a multiplicity of rival traditions, each with its own characteristic modes of rational justification internal to it, then that very fact entails that no one tradition can offer those outside it good reasons for excluding the theses of its rivals." MacIntyre 1988, 352.

[60] Even noticing the fact that two traditions are each other's rivals presupposes some amount of understanding each other. MacIntyre 1988, 370.

[61] MacIntyre 1988, 372.

learned like one learns one's mother tongue. Essential in this process of learning is that one learns to understand an alien tradition from inside as if it were one's initial tradition.[62]

After one has learned to understand the language, practices, and institutions of another tradition, the second stage in the process of translation follows. Now the task is to express the other tradition in terms of one's own first language, that is to say, in terms of the language, practices and institutions of one's initial tradition. Accomplishing this requires, however, that the linguistic and conceptual capacity of one's own tradition can be enlarged. The translation has been successfully carried through if a participant of the rival tradition can identify her own tradition in that translation.[63] Two things take place during the process. First, the translation establishes an *understanding* between one's own tradition and its rival; second and even more importantly, a successfully completed translation proves that one's own tradition has survived an *epistemological crisis* caused by the encounter with the alien tradition.[64]

What else does surviving an epistemological crisis mean for a tradition? Outliving such critical situations encapacitates the adherents of a tradition to reformulate its history in a new, more perspicacious way. It means that the tradition has developed concepts explicating the continuities according to which the tradition has survived as one and the same tradition. In this way, the tradition receives a specific identity or a histor-

---

[62] MACINTYRE 1988, 374.

[63] MacIntyre does not seem to notice that his reasoning becomes circular here. Let us suppose that a translation has been made from the tradition *A* into the language of the tradition *B*. The translation is correct if a participant from the tradition *A* can understand it. But how could she do this without any knowledge of the language etc. of the tradition *B*? Would she not first have to learn the tradition *B* and translate its central concepts into the language of the tradition *A* before she could determine whether the original translation from *A* to *B* is correct or not?

[64] MACINTYRE 1988, 374–375, 385, 387–388. WESTON (1991, 403) and MARKHAM (1991, 262–267) point out that MacIntyre's method for evaluating divergent traditions against each other cannot actually succeed. MacIntyre's programme presupposes that one is committed to a concept of rationality which allows for the possibility of one's own tradition to err. For a committed member of any strong, say, religious tradition this is an impossible option; one of the basic beliefs of such a tradition is that it contains the whole truth and nothing but the truth.

ically constituted self-understanding, which becomes an integral part of how the members of the tradition understand themselves and their lives. Furthermore, surviving an epistemological crisis highlights the structure of epistemological justification within that tradition. All in all, a tradition which has survived an epistemological crisis successfully both covers a wider range of phenomena than its rivals and is able to do this according to its own standards of justification.[65] In contrast, a tradition that fails in the task of translation perishes mainly because it cannot pass the tests of justification, which the standards of that tradition set. The possibility of such failure shows that the premises of the *relativist challenge* are mistaken. The relativists, namely, maintain that, although mutually incomparable, each tradition can always be vindicated by its own standards in a way that makes failing according to its own criteria of rationality impossible.[66]

Despite the fact that traditions possess and apply different sets of criteria for rationality, they can confront each other and one tradition can defeat another. Therefore, it is in the adequacy or inadequacy of the response which a tradition can bring forth in the face of an epistemological crisis that it is vindicated, or that it is defeated.[67] This view differs greatly from the relativist conviction: a relativist observer typically conceives of herself as an outsider, considering different traditions from an outside, neutral perspective. The idea of an impartial observer is, nevertheless, misconceived. According to MacIntyre, a neutral perspective is a

---

[65] MacIntyre 1988 363.

[66] "Every tradition, whether it recognizes the fact or not, confronts the possibility that at some future time it will fall into a state of epistemological crisis, recognizable as such by its own standards of rational justification, which have themselves been vindicated up to that time as the best to emerge from the history of that particular tradition. All attempts to deploy the imaginative and inventive resources which the adherents of the tradition can provide may founder, either merely by doing nothing to remedy the condition of sterility and incoherence into which the enquiry has fallen or by also revealing or creating new problems, and revealing new flaws and new limitations. Time may elapse, and no further resources or solutions emerge. That particular tradition's claims to truth can at some point in this process no longer be sustained. And this by itself is enough to show that if part of the relativist's thesis is that each tradition, since it provides its own standards of rational justification, must always be vindicated in the light of those standards, then on this at least the relativist is mistaken." MacIntyre 1988, 364.

[67] MacIntyre 1988, 363–366.

conceptual impossibility; the notion of understanding presupposes understanding from a certain point of view. To be outside all traditions means that one lacks any sufficient rational resources for inquiry, and hence, for understanding. For this reason, it is nonsensical to state anything outside all traditions. Furthermore, the relativist fails to see that even the possibility of issuing the relativist challenge requires a certain point of view, a point of view internal to a particular tradition.[68]

A critique, analogous to that directed towards relativism, can be applied to the perspectivist challenge, with its view that all traditions represent different perspectives on the same things or display diverging aspects of reality, but that no position can rise above the others and claim to embody the truth. Like the relativist, the perspectivist fails to notice that one cannot adopt a neutral, outside position in relation to traditions: the fact that one states something of reality already implies that one has adopted a specific point of view. Additionally, perspectivists do not see that the possibility of translating a tradition into the language of another tradition makes it feasible that one tradition not only defeats another but that it also fails according to its own standards of rationality, consistency, etc.. The perspectivist challenge is powerless against the tradition-bound form of inquiry.[69]

In contrast to the relativist's and perspectivist's claim, it is legitimate, according to MacIntyre's view, to speak about the agreements of a surviving tradition as *true*, as something that correspond with reality irrespective of time and place in a way adequate to the object of the asserted propositions; and not just, in the relativist mode, state that there is "warranted assertability" in relation to a tradition.[70] This means that MacIn-

[68] MACINTYRE 1988, 367.

[69] "The perspectivist's failure is complementary to the relativist's. Like the relativist the perspectivist is committed to maintaining that no claim to truth made in the name of any one competing tradition could defeat the claims to truth made in the name of its rivals. And this we have already seen to be a mistake, a mistake which commonly arises because the perspectivist foists on to the defenders of traditions some conception of truth other than that which is theirs, perhaps a Cartesian or an Hegelian conception of truth or perhaps one which assimilates truth to warranted assertibility." MACINTYRE 1988, 367.

[70] MACINTYRE 1988, 363–364.

tyre's tradition-constituted and tradition-constitutive mode of enquiry is an attempt to re-enforce the concept of truth into the scheme of enquiry. On the basis of what has been said above, we can distinguish two different aspects to truth. From the perspective of a tradition that has survived its past and present epistemological crises its central pronouncements represent truth in an absolute, or realist sense. This truth has been formulated in the language and with the concepts of the tradition, and it has been vindicated in terms of the justificational criteria of that tradition. Despite this, it would be wrong to say that this view of reality is just relational to the particular tradition: MacIntyre stresses that to say something of anything necessarily means adopting a point of view; hence it is nonsensical to speak about an absolute, or non-relative truth which has no relation to any particular tradition.

We can consider the concept of truth from yet another angle which MacIntyre does not, however, explicate. As has been noticed, every tradition confronts the possibility of falling into an epistemological crisis so deep that the tradition fails in its own terms. From this point of view, the present truth within a tradition may not be lasting, and in this sense, the truth of a tradition is always relational to the further and future stages and forms of that tradition.

Can we extend MacIntyre's model to rationality, and say that one tradition is more rational than another, even if we cannot appeal to any criteria of rationality external to those traditions? We can, maintains MacIntyre, and the most rational thing to do is to adopt the tradition which has survived the challenges presented by its rival traditions. That tradition has been able to modify and enlarge its conceptual potentialities and has adapted itself to the historical reality in which the questions posed by the contending traditions, and conflicts springing from the diverse historical heritage of the tradition itself need answering. A rational agent who seeks a rational, tradition-bound way of inquiry ought to choose the tradition capable of overcoming the epistemological crises caused by the confrontation of questions, problems and contradicting answers from two or more different traditions. A tradition which has passed through an epistemological crisis can successfully cover the wid-

est range of problems, coming both from inside and from outside that tradition. It can give the most satisfactory answers to a rational agent who is in search of a solution to such problems.[71]

We have now examined the constituents of MacIntyre's revision programme. It is time to consider how the different parts of his theory connect with each other and what kind of a concept of a person emerges from the theory. On the basis of the analysis we can say that the identity of a tradition is analogous to that of a person in MacIntyre's theory. In both cases, when we are speaking of a tradition or of a person, the identity is constituted by a historical past consisting in a series of contingent occurrences. But this is not all: the history is a narrative. There is a certain underlying concept which constitutes both the self-interpretation of a tradition and a person's self-identity, or in other words, what it means to a particular tradition or to a person to have just a certain kind of a past. In both, there are specific commitments, practices and beliefs which structure the past and give it a distinct form. The continuity of a tradition, or that of a person, is never just a sum of certain external characteristics but involves a specific interpretation of the past as constitutive of the present form and content, and of the identity of the particular tradition or person. This conceptual scheme, in which the past of a person as well as that of a tradition are evaluated forms, what could be called, an *essence* of the person or the tradition.

MacIntyre's theory has been accused of extreme conservatism for it presents no criteria outside the prevailing form of a tradition which

[71] MACINTYRE 1988, 388. See also MACINTYRE 1988, 393–395 where MACINTYRE speaks especially to readers who are dissatisfied with the modes of enquiry they have contested so far, and who are in search of rational means to decide between different modes of enquiry. MacIntyre's conception of the most rational tradition and the process of epistemological "testing" and choice leading to it resemble Brandt's procedure of cognitive psychotherapy, although the coverage of MacIntyre's conception is much larger, and Brandt deals mostly with moral phenomena; see "The method of cognitive psychotherapy" on page 28. In both theories the need to change the frame of one's thinking arises from epistemic dissonance. The main difference lies in the source of criteria for determining the rationality of one alternative against the other. Brandt offers a list of ready criteria for doing this, whereas MacIntyre leaves the question of criteria open for conceptual innovation; the theory which consistently covers more phenomena than any other is held to be the best.

could be used for criticizing it.[72] On the basis of our analysis we can, however, say that this criticism along these lines is not wholly justified. What I have here called the essence of a tradition or a person could indeed be used as a critical corrective and evaluative criterion. The essence represents a form of understanding and a scheme of interpretation which is constitutive of the identity of a person or of a community, and this can be applied as an evaluative criterion for prevailing institutions and practices, and for persons. It is, however, true that MacIntyre's model does not necessarily invite any profound criticism against prevailing social and political traditions.

As we have noticed above,[73] it is possible for a tradition to become extinguished, to cease to exist. If this happens, we can no longer identify a certain tradition, once distinguishable from others, as a distinct, independent tradition. Accordingly, there will no longer be a group of people who identify themselves as bearers and continuers of that tradition. It is feasible, against this background, to consider the possibility of a person ceasing to exist. Could a person lose her identity, so that she could no longer identify her present self with the person that she once was? Could some Parfit-like development take place in a MacIntyrean person?

MacIntyre describes the post-Enlightenment conception of personal identity as the modern liberal self: it is a shattered being, and because it has been disconnected from its past it can no longer have a lasting identity. But it seems that MacIntyre's own conception of a person faces a similar danger. MacIntyre's concept of a person and of personal identity do not differ from that of a tradition, and of the identity of a tradition. A tradition ceases to exist if its basic commitments are abandoned. Certain features of a moribund tradition may continue their life as a part of

---

[72] E.g., Nussbaum (1992, 10) firmly indicts MacIntyre's view for cultural and political conservativism, an affinity to ethical traditions which endorse a hierarchy of authoritarian and subordinate roles, fixed social positions giving little if any possibilities for criticism and change. The basic reason for MacIntyre's conservativism is that he abandons all the ideas of the Enlightenment without seeing that it represents a tradition which gives a legitimate and meaningful role to reflection in ethical thinking, and a feasible concept of universality and of impartiality. See also Almond 1990, 100 and Pence 1984, 284.

[73] See page 229.

another tradition, but mere historical connection or continuity does not constitute the identity of a tradition.[74]

In MacIntyre's view, there must always exist commitments and continuities which are internally, not solely externally, definable. This means that a tradition ceases to exist if there is no community which identifies itself through the basic commitments and interpretative schemes of that tradition. By analogy, it is possible to imagine that a human being goes through a process similar to an epistemological crisis, and that the person fails to "survive" this crisis.[75] This means that the person could not overcome the difficulty in terms of her own historically and socially determined commitments; her adopted self-understanding could not help her, and she could not make sense of her life by referring to her interpretative scheme of reality. It is possible to think that the crisis could be so deep that the person would have to abandon her old identity completely and replace it with a new identity including new kinds of intepretative schemes, to the degree that she could no longer identify herself with the past of her former self, not even understand this previous person. In this case, the only link between the person's past and present selves would be a certain physical, external continuity.

Such external continuity cannot, however, be enough to constitute personal identity in the stronger, internal and interpretative sense MacIntyre wants to apply: even if the past and the present selves of the person have gone through the same historical phases, the interpretation they give/gave to them can be completely dissimilar to each other, because of their different commitments. MacIntyre's concept of identity is a strong one, whereas all external criteria represent only weak versions of identity. We can then say that MacIntyre's concept of a person allows for the possibility that a person ceases to be although she is not dead yet, and is fully conscious, and another person will emerge, one who shares her external past with the "first" person. The MacIntyrean person can, thus, face the same kind of a destiny as the post-Enlightenment disconnected selves

[74] See, e.g., MacIntyre 1988, 384.

[75] For examples of this, see MacIntyre 1988, 394–395.

which MacIntyre so strongly criticizes. This possibility gives us a reason to suspect the rationality of MacIntyre's whole programme: why return to the philosophical ideas which preceded the Enlightenment? Were they not just the schemes of thought which historically led to those difficulties MacIntyre so strongly criticizes? Would it not be better to try to develop conceptual novelties for solving these problems?

MacIntyre's problem seems to be that in order to introduce a strong concept of identity, he has to apply certain internal criteria for identification. In the case of traditions this is not problematic, because it does not matter whether a tradition dies out or not: the application of internal characteristics as criteria for identity does not lead to difficulties. But when applied to persons, criteria that are purely internal become problematic. MacIntyre seems to think that it is not possible to lose one's identity in this radical sense, but his own theory is defective because a total loss of identity remains a theoretical possibility. To avoid the problem MacIntyre would have to supplement his conception of personal identity with external criteria.

I will end the present section by presenting a discussion from Paul Ricœur.[76] His ideas resemble those of MacIntyre's for he also employs a narrative conception of a person, but in his theory the criteria of personal identity are not solely internal, but also external. Ricœur distinguishes two types of personal identity, an *identity*$_{\text{idem}}$ and an *identity*$_{\text{ipse}}$. They correspond with two approaches to human personhood which are the *third-person* and the *first-person* points of view. Furthermore, these two aspects have an analogy in the human intellectual approach: analytical thinking which regards the agent "from outside" explicates the conditions of identity$_{\text{idem}}$, whereas (subjective) reflection "from inside" clarify the nature of identity$_{\text{ipse}}$. Identity$_{\text{idem}}$ represents the idea of *sameness* in contrast to identity$_{\text{ipse}}$, which designates identity as selfhood, i.e., the fact that a person recognizes herself as herself. The two approaches to personal identity are further distinguished by their specific criteria. The cri-

[76] Paul Ricœur, *Soi-même comme un autre*. Seuil, Paris, 1990. I take up only that part of Ricœur's discussion which is relevant for the present theme, and leave out his idea of the self as other.

teria for identity$_{idem}$ are external, while identity$_{ipse}$ is determined along internal criteria. Identity$_{idem}$ and identity$_{ipse}$ are, however, linked in the concept of *character* which is both something externally recognizable and internally knowable, and which allows both permanence and change without loss of identity.

Ricœur also employs the notion of *narrative identity* as a link between the two aspects to identity. By this he means that a person (or a community) has values, norms, ideals, etc., which she uses for self-identification, i.e., for recognizing herself as a part of the personal and communal history formed through socially constituted meanings and values that impinge on her. We can express the thought by saying that there is a conceptual framework upon which a person's self-identification as identity$_{ipse}$ is based. This self-understanding is, however, also a part of the identity$_{idem}$, for we cannot give a description of a person without referring to her narrative identity and thereby also to her historically formed self-understanding. Ricœur's two-sided concept of personal identity prevents the kind of radical loss of identity which is possible in MacIntyre's theory.

3.5. The self and its sources — Charles Taylor's theory

Charles Taylor is not a virtue ethicist.[1] Examining his ideas in this connection is, however, meaningful for various reasons. His criticism of the modern understanding of morality and his interpretation of the history of ideas resemble MacIntyre's views. Like MacIntyre, Taylor uses material from the history of philosophy for explicating current philosophical models of thought; the sources of modern thinking lie in history, and the present cannot be understood without understanding the past. Taylor also offers a conceptual scheme for examining the nature of morality which is worth studying in the present context. This section will begin with some points from Taylor's criticism against the conceptual framework of both modern humanities and moral philosophy. The analysis then moves to Taylor's suggestion for an alternative way of understanding human life and ethics. This forms a basis for explicating Taylor's conception of a person. The chapter ends with a section in which Taylor's and MacIntyre's views are compared with each other.

Like the rest of the humanities, modern moral philosophy has adopted its conceptual scheme from the natural sciences. Although this is understandable in the light of the great progress physical sciences have made during the past centuries, it is, nevertheless, a mistake. The way of thinking and the understanding of reality which natural sciences represent does not permit an explanation of human action and the kinds of phenomena characteristic of human intentionality. The conceptual apparatus of the natural sciences does not match with the reality of human behaviour, and theories which have been structured according to this ideal fail to explicate what is essential in human life.

---

[1] Charles TAYLOR, *Sources of the Self: Making of the Modern Identity*. Cambridge University Press, Cambridge, 1989; *Human Agency and Language. Philosophical Papers 1*.; *Philosophy and the Human Sciences. Philosophical Papers 2*. Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney, 1990.

Modern moral philosophy is obsessed with what Taylor calls a *naturalist temper*. By this Taylor means that the model of explanation used in moral theories derives from the physical sciences, and the ideal moral theory is conceived of as providing a "scientific" explication of moral phenomena.[2] Two features characterize this approach. First, human behaviour and human institutions are given an explanation in terms of an impersonal, "neutral" and "scientific" language which describes the phenomena externally and detaches them from culturally determined frameworks of meaning.[3] Second, ethical thinking has become *procedural,* in the sense that it focuses on procedures on the basis of which the outcome of the ethically correct action can then be determined. This becomes manifest in both deontological and utilitarian theories and in their concentration on the rightness of actions instead of the moral quality of agents.[4]

This development has had serious consequences for moral philosophy. The externalist mode of speech has made it impossible to speak of anything as having meaning.[5] Modern ethics has made morality a sepa-

---

[2] "One of the defining characteristics of naturalism, as I am using the term, is the belief that we ought to understand human beings in terms continuous with the sciences of extra-human nature. Just as these last have progressed by turning away from anthropocentric language, by excluding descriptions which bear on the significance of things for us, in favour of 'absolute' ones, so human affairs ought to be maximally described in external, non-culture-bound terms. Thinkers of a naturalist temper, when considering ethics, naturally tend to think in terms of action. This temper has helped contribute to the dominance of moral theories of obligatory action in our intellectual culture." TAYLOR 1989, 80–81.

[3] TAYLOR 1989, 3–4, 9–10, 14.

[4] "Much contemporary moral philosophy, particularly but not only in the English-speaking world has [...] tended to focus on what it is right to do rather than on what it is good to be, on defining the content of obligation rather than the nature of the good life; and it has no conceptual place left for a notion of the good as the object of our love or allegiance [...]" TAYLOR 1989, 3. "Moral philosophies so understood [as defining the content of obligation rather than the nature of the good life] are philosophies of obligatory action. The central task of moral philosophy is to account for what generates the obligations that hold for us. A satisfactory moral theory is generally thought to be one that defines some criterion or procedure which will allow us to derive all and only the things we are obliged to do." Taylor 1989, 79. Taylor maintains that his criticism hits both the deontological and the utilitarian strands of modern moral philosophy. TAYLOR 1989, 88–89.

rate phenomenon which is disconnected from ontological commitments and evaluative language; meaning-giving frameworks have been expelled to the private and to the personal, which lie outside scientific discourse.[6] Likewise, the subject of modern moral philosophy is an individual, a solitarily conceived self who lives detached from historical communities of human beings, and culturally determined modes of understanding.[7]

Modern moral philosophy has become respectably "scientific" but the price it has had to pay for this is high: it has become incapable of performing its own task of understanding human life.[8] The development can also be seen from another angle. Abandoning the language of commitment and meaning has made modern moral philosophy blind to the metaphysical commitments and ontological premises which, as a matter of fact, direct its own modes of thinking and understanding. [9]

If the situation is as bad as Taylor claims it to be, what should be done? What could correct the state of affairs in moral philosophy? To

---

[5] In Taylor's view one of the main reasons for the strong sense of confusion inside modern ethical theories, as well as the problem of meaninglessness in people's lives, are connected together with the fact that all traditional frameworks are discredited to the status of personal predilection, and are, hence, dispelled from the legitimate sphere of ethical study. We have lost the context within which the question of meaning could have its place. TAYLOR 1989, 9–10, 16–17.

[6] In ethics, Taylor maintains, we deal with questions concerning how we could best understand human life and human action. This is not possible without a conception of the self. But to study the self has become difficult; in modern science the legitimate method of scientific research is that of the natural sciences. This means that the object of study must be taken "absolutely" not in its meaning for us or for any other subject, but as it is on its own, i.e., objectively. The object of study must be seen as independent of any descriptions or interpretations offered of it, the object can in principle be captured in explicit description; and this description is without reference to its surroundings. To speak about a self in these terms, however, destroys just what is essential for the subject matter, that is, the language of self-interpretation. TAYLOR 1989, 33–34; TAYLOR 1990a, 1–2. Taylor's criticism is similar to that of MacIntyre's, see page 206.

[7] "One of the most negative of these features [of modern identity] is atomism. The disengaged identity and its attendant notion of freedom tend to generate an understanding of the individual as metaphysically independent of society." TAYLOR 1990a, 8. About the evolvement and characteristics of this atomism, see TAYLOR 1990b, 187–210. According to Taylor, it is noteworthy that characterizations of the modern moral agent are devoid of any notions that would link it together with, or bind it to a particular, historic community with given webs of birth and history. See TAYLOR 1989, 11–12, 37, 84.

[8] TAYLOR 1989, 57–58, 98.

find an answer to this, it is necessary to examine Taylor's alternative approach to ethics. As before in this study, attention is paid not only to Taylor's normative views but also to his method and his central theoretical concepts. Taylor's approach is historical, he analyzes modern philosophy by telling the story of the sources of the modern self.[10] Is this simply a mode of presentation, or does it have deeper, methodological significance? Does the historical approach have an impact on the conception of a person underlying the theory? As Taylor's criticism against modern moral theories shows, he regards the scope they give to morality as too narrow. For him, the task of ethics is not limited to defining the content of obligation, or for establishing the criteria for right actions. Ethics must, rather, examine questions which deal with the problem of what makes a life meaningful, or explore what it is to be a human agent, a person, or a self.[11] Underlying Taylor's understanding of ethics there is a theoretical framework which I will now investigate further.

It is a fact, Taylor maintains, that all people, always and everywhere, have moral intuitions or a sense of the ethical. This moral sense can be characterized, in the most general sense, as having three constituents. First, all people have a sense of respect for and of obligations to others. This forms a cluster of powerful basic moral demands that we feel to be directed at ourselves. The second constituent is our understanding of

[9] "But I think that for all the diversity of these reductionisms they form a family nonetheless. What they have in common is a certain metaphysical motivation. [...] In fact the motivation is many-faceted, but one way of defining it is via the paradigm status accorded to the natural sciences as the models for the sciences of man. In a certain sense of the term, this family of theories shares an allegiance to 'naturalism', by which I mean not just the view that man can be seen as part of nature [...] but that the nature of which he is a part is to be understood according to the canons which emerged in the seventeenth-century revolution in natural science." TAYLOR 1990a, 2. See also TAYLOR 1990a, 46.

[10] "But I don't think we can grasp this richness and complexity [of modernity] unless we see how the modern understanding of the self developed out of earlier pictures of human identity. This book attempts to define the modern identity in describing its genesis. [...] I try to set out in the concluding chapter what flows from this story of the emerging modern identity. [...] Understanding modernity aright is an exercise in retrieval." TAYLOR 1989, ix–x.

[11] TAYLOR 1989, 3–4, 14–15. For Taylor "human agent", "person" and "self" all represent the same concept.

what makes a life worth living, and the third is our sense of dignity, or our view of ourselves as the object of moral respect on the part of others.[12] The norms governing the relationships between people differ extensively from community to community and from time to time, but it is still true that we all have moral intuitions as strong as instincts. This is a factual proposition which applies to all human beings.[13] Furthermore, moral intuitions seem to involve claims, implicit or explicit, about the nature and the status of human beings. The sense that human beings are fit objects of respect and that their life and integrity is not to be infringed corresponds with our moral intuitions.[14]

Now, Taylor employs the universal phenomenon of a moral sense of respect and an according behaviour as the basis for his concept of ethics. In his view, moral sentiments and institutions are not intelligible unless we presuppose that they are supported by a certain ontological framework. Accordingly, a moral reaction is always also an affirmation of a given ontology concerning the concept of the human.[15] And this ontology constitutes the conceptual framework, which is necessary for understanding human agency and life. It is our "mode of access to the world in which ontological claims are discernible and can be rationally argued about and sifted."[16]

What is this ontological framework and how does it relate to moral phenomena? The conception of a moral ontology is a part of Taylor's

---

[12] Taylor 1989, 4, 15.

[13] "Virtually everyone feels these demands, and they have been and are acknowledged in all human societies. Of course the scope of the demand notoriously varies[...] But they all feel these demands laid on them by some class of persons, [...] We are dealing here with moral intuitions which are uncommonly deep, powerful, and universal. They are so deep that we are tempted to think of them as rooted in instinct, in contrast to other moral reactions which seem very much the consequence of upbringing and education." Taylor 1989, 4–5.

[14] "So our moral reactions in this domain have two facets, as it were. On one side, they are almost like instincts, comparable to our love of sweet things, or our aversion to nauseous substances, or our fear of falling; on the other, they seem to involve claims, implicit or explicit, about the nature and status of human beings." Taylor 1989, 5. See also Taylor 1989, 25.

[15] Taylor 1989, 6, 7.

[16] Taylor 1989, 8.

larger, hermeneutical scheme for studying human life and action. Human beings are essentially "self-interpreting animals". This means that all human experiences and emotions are indivisibly attached to modes of speech and action which give meaning and importance to these experiences and emotions. Consequently, even a seemingly immediate desire or feeling is partially constituted by the linguistic form and vocabulary which are used for referring to that desire or feeling. Furthermore, many of our desires and feelings are subject-referring: to feel shame, pride or love, includes a description of the feeling subject as rational being or moral being, or the like. Such descriptions include implicit criteria for a class of, say, rational or moral, beings which are capable of these emotions. Articulating the meaning of such feelings and desires is to incorporate a sense of what is important for us in our lives as subjects. This sense is the key to our self-understanding as human beings.[17]

What then are the implications of Taylor's idea? For him there is no "purely" empirical access to human reality, but everything from feeling and desiring to intentional action takes place within an interpretative scheme, or an ontological framework. Thus, "a fully competent human agent not only has some understanding of himself but is partly constituted by this understanding".[18] The two are indistinguishable: there are no experiences and the like outside an understanding of what it is to have that experience. And the interpretation does not exist independently of human reality but only as its part. Human reality is an interpreted reality and the world has normative and evaluative aspects. This means that human reality is structured by an ontological framework, and that one cannot understand this reality without and irrespective of such frameworks. This is the conceptual starting point of Taylor's theory. Does it have an impact on his concept of a person?

Our moral ontology is often implicit, but when we articulate it we do it in the form of a framework which explicates what makes sense of our

[17] TAYLOR 1990a, 56–61.

[18] TAYLOR 1990a, 3.

moral responses.[19] Although actual people's moral senses and their ontological frameworks vary culturally and historically, this does not mean that frameworks are simply arbitrary interpretations of reality. On the contrary, frameworks are constitutive of human agency, and they create the only basis for a meaningful language of human personhood and selfhood.[20] Against this background, it is clear why the modern scientific approach deriving from natural sciences is so misguided when it is applied to the study of human phenomena. If we want to understand human life we must use the language and modes of explanation which are appropriate for the object of study. If it is the subjective, interpretative mode of speech then we simply have to employ it.[21]

Human existence is historically and culturally determined, because ontological frameworks are always tied to a specific historical and cultural situation. There are no modes of moral thinking distinguishable from the way in which people make sense of their particular lives, and through which they conceive themselves as human agents living at present.[22] But does this mean that we cannot explicate one single concept of a person in Taylor's theory when that "person" is a notion which must

---

[19] Taylor 1989, 19. See also Taylor 1989, 26.

[20] Taylor 1989, 26–27. For Taylor moral horizons, or frameworks of qualitative discriminations are closely linked together with the question of human identity: "My identity is defined by the commitments and identifications which provide the frame or horizon within which I can try to determine from case to case what is good, or valuable, or what ought to be done, or what I endorse or oppose. In other words, it is the horizon within which I am capable of taking a stand." Taylor 1989, 27. "We can say therefore that our self-interpretations are partly constitutive of our experience. For an altered description of our motivation can be inseparable from a change in this motivation. But to assert this connection is not to put forward a causal hypothesis: it is not to say that we alter our descriptions then *as a result* our experience of our predicament alters. Rather it is that certain modes of experience are not possible without certain self-descriptions." Taylor 1990a, 37. See also Taylor 1989, 34, 1990a, 34–35.

[21] Taylor calls this principle the BA (best account) principle; we must use the best mode of explanation open to us, and we "cannot just leap outside of these terms altogether, on the grounds that their logic doesn't fit some model of "science" and that we know a priori that human beings must be explicable in this "science". This begs the question. How can we ever know that humans can be explained by any scientific theory *until* we actually explain how they live their lives in its terms?" Taylor 1989, 58.

[22] Taylor 1989, 35–36, 56.

be defined only against the particular background of a specific cultural setting? Despite the importance of historical particularity, it is possible to define the concept of a person here by using Taylor's conceptual starting point as a basis. Even though the *content* of each interpretative framework is specific and particular, it is universal as an interpretative *form*. The "essence" of human personhood is an existence which actualizes as experiences, actions, feelings, desires, etc., within an interpretative meaning-giving framework. Interestingly here, Taylor's conception of a person is analogous to his description of morality. The content of morality changes historically, but what remains universally unaltered is the form: an ontological framework within which people can find things valuable and meaningful.

As we noticed before in MacIntyre, a theory in which historical particularity receives a central status runs the risk of becoming radically relativist.[23] How does Taylor avoid this threat? How should the problem of two incompatible and contradicting frameworks be solved?

In case two alternative ontological frameworks conflict, we have to *reason in transitions*. This is a mode of practical reasoning in which the aim is not to establish something absolutely, but to discover some position's provisional superiority in relation to some other. Being confronted with two contradicting claims, we can say that one of them is well founded if we can show that the *move* from position A to position B consitutes an epistemic gain. This means showing , for instance, that we get "from A to B by identifying and resolving a contradiction in A or a confusion which A relied on, or by acknowledging the importance of some factor which A screened out, or something of the sort."[24]

In the process of understanding oneself as a human being one's reasoning is transitional in this sense.[25] There cannot be any criteria independent of a person's particularly determined perspective.[26] Accordingly, if one must defend one's ethical point of view, one must do this by telling a story of the genesis of this particular moral position. One then expli-

[23] See page 227.

[24] TAYLOR 1989, 72.

cates the moral ontology and the meaning-giving framework which underlie the view.[27] The most reliable moral view is the "one that is grounded on our strongest intuitions, where these have successfully met the challenge of proposed transitions away from them."[28]

There are features in Taylor's theory which resemble MacIntyre's views. First, the concept of a framework in Taylor's theory corresponds

---

[25] Taylor also speaks about "radical re-evaluation". By this he means challenging one's most fundamental evaluations that touch one's identity and that provide the horizon for the other evaluations one makes. Such evaluation is especially difficult because one cannot use a metalanguage for assessing rival self-interpretations. TAYLOR 1990a, 40. Taylor compares this task with that of a philosopher: "We start off with a question, which we know tobe badly formed at the outset. We hope that in struggling with it, we shall find that its terms are transformed, so that in the end we will answer a question which we could not properly conceive at the beginning." The same method applies to radical re-evaluation. TAYLOR 1990a, 41.

[26] The criteria establishing "an epistemic gain" are partly logical, partly what could be called rational in a wider sense. A detected self-contradiction in any theory is a clear reason for ruling it out as an alternative for self-understanding. If B covers a wider space of "phenomena" than A, it seems to qualify as the winner of the two. TAYLOR 1989, 72. In an essay entitled "Rationality" (TAYLOR 1990b, 134–151) Taylor discusses the problem of comparing two incommensurable cultural systems with their own distinct and incommensurable criteria of rationality, using the Western theoretical culture and an atheoretical culture as examples. He tries to establish some transcultural criteria for regarding the one, in this case the theoretical culture, as more rational than the other. Taylor maintains that no human cultural systems are completely alien to each other, they all offer articulations and select different features of the world and human action in some perspicuous order. In comparing two different ways of laying out an order, one culture can be said to propose a higher, or fuller, or more effective rationality, if the order it offers is more perspicuous than the one adopted by the other culture. Even if the aims of the compared cultures are different and conceived of differently, there are always common goals in every culture. Accordingly, the theoretical culture of the West with its view of rationality is, at least in some respects, higher than an atheoretical culture: the theoretical culture proves to be more effective, and hence more rational in the face of its technological achievements. Taylor justifies his conclusion by maintaining that irrespective of the way a human community understands its life, it always attempts to develop technologies with the help of which it will be possible to use nature for the good of the community. Western theoretical culture provides a better means for building up effective technology for this purpose than any atheoretical culture. So, at least in this respect, it is a more rational model for structuring reality and human life in it. Even if there is cultural plurality, and a plurality of concepts of rationality connected with the way human activities are seen, it is possible to make comparisons of superiority between different cultural systems. It seems, though, that despite this possibility, there are no absolute, nonrelational criteria for establishing rationality; the only method for judging rational superiority is a comparative one.

with the notion of a tradition in MacIntyre's theory. A framework plays the same epistemological role as a tradition: no one can adopt a position, or claim to know something outside and irrespective of some framework. Taylor's 'moral ontology' has the same significance for understanding human life as has MacIntyre's idea of the necessity of a particular, historically determined tradition. In both theories, this central concept provides a basis for the concept of a person. Furthermore, Taylor speaks of "reasoning in transitions" whereas MacIntyre employs the term "solving an epistemological crisis". Both Taylor and MacIntyre acknowledge the possibility of translatability between different, mutually incommensurable traditions, and they both insist that a controversy between two such rivals can be solved rationally. They further maintain that we can, and must speak of ethical truth in realistic terms, but that this truth is historically and particularly determined.

[27] TAYLOR 1989, 72–73. This is also Taylor's aim: he tells the genesis of the modern moral theories to bring to light their implicit moral ontologies and moral motivations.

[28] TAYLOR 1989, 75.

3.6. The concept of a person in virtue theories

It is time to sum up some characteristics that have held our attention as we have scrutinized virtue theories. As was suggested in the beginning of this section, a mutual motivator of virtue theories is a critique directed towards dominant approaches in moral philosophy. The main point of this criticism is that current moral philosophy has given up all conceptual means for attributing moral properties to reality. All the writers I have examined in this section are opposed to the sharp distinction made between factual and evaluative statements. Many of them also deny that there is actually any dissimilarity between moral and other evaluative propositions but that these two are linked together. In Philippa Foot's model the link between factual and evaluative propositions is to be found in the moral terms themselves, whereas Wallace reckons that the meaning of ethical statements arise from the mutually shared conventions of a community. Nussbaum and MacIntyre intentionally seek a methodological alternative in the classical Aristotelian tradition, although the ethical models they deduce from these premises are fundamentally different.

The theories we have examined in this chapter share a critical attitude towards many conventional models of moral philosophy. Although the alternatives they suggest for replacing the established forms of utilitarianism and contractarianism differ from each other, there are still some features they have in common. The theories of this chapter represent a view according to which there is a plurality of ethical values. This is a plurality of goods which cannot be subsumed to one super-category of moral worth. Each of these goods is valuable and irreplacable in itself and the deprivation or loss of one good cannot genuinely be compensated by another, different good. Another typical feature in these theories is their emphasis on the institution of morality always taking its actual form in a historically determined social community. What combines these two features is the concept of a moral virtue. There is a plurality of moral virtues which are as little reducible to one single virtue as

the goods of human life are contractable to one single good. Against the background of a good human life, the practising of these virtues is both a means and an end in itself. Virtues are a means in the sense that they sustain practices which support a good life in a community. Additionally, they are an end in themselves since they are constitutive of a form of life which makes human flourishing possible. But can we, in addition to such common features, find a concept of a person which is common to all these theories?

On the basis of the preceding analysis, we are justified in sketching a concept of a person for the ethics of virtue. As we have already noticed in connection with utilitarianism and social contract theories, the definition of the morally relevant gives us the central characteristics of the moral person. What the moral theorist must pay attention among the phenomena of human life is the plurality of moral values, the real meaning of which cannot be understood unless they are seen as a part of the life of a historically determined community, its ways of thought and its socially determined practices. To be understood correctly, morality must be seen against the background of a broad framework, against that of a tradition where communities are concerned, and against that of an individual life where a human person is at issue. The same features which define the morally relevant also characterize the moral person. The plurality of values as a starting point has the effect that the concept of a person is not formal as in the two sets of theories we have examined earlier; there is much more "substance" in the person of virtue ethics. The perspective in which the person is now seen is her whole life, conceived through the understanding the person has of herself, of her past and of her future perspectives. The person of virtue ethics is really the story of her life, understood and structured by a socially shared framework of practices and beliefs. We can, thus, characterize this as a narrative concept of a person.

According to the ethics of virtue, one does not comprehend the real nature of morality if one regards ethical questions through some single principle or value. The same applies to the concept of a person: it is a rich notion the correct understanding of which presupposes the context

of a detailed story. Moreover, the story is not a given or a ready-made report but an on-going narrative which changes as the person's understanding of the phenomena of life in general and her view of good human life in particular acquire new forms during the course of her life. This means that being a moral person is essentially something other than that which is indicated by the utilitarian or the contractarian theories. It is not being a desiring agent or adopting a specific role of a moral actor, but extensive and continuous reflection on what it involves to live as a human being among other human beings.

When examining the utilitarian moral person, I characterized the utilitarian point of view towards morality as moral imperialism, as a view in which the moral point of view covers all other perspectives of human life. As the preceding study has showed, the narrative conception of a person introduces a very extensive view of human life, too. Does this conception represent a form of moral imperialism comparable with the utilitarian approach? The answer to this question is a qualified "yes". Utilitarianism offers a model which tends to mould all aspects of human life to fit the utilitarian moral approach, and the virtue theories do the same albeit in a different manner. These theories accept as their starting point a multiplicity of goods which represent different perspectives on human life. Adopting a narrative conception of life does not squeeze everything into one scheme of thought as in utilitarianism but it still represents an extensive overarching perspective that replaces all other viewpoints. A person cannot have a life outside the narrative of her life.

4. *The concept of a moral person*

In the beginning of this study I accepted some presuppositions to direct the examination of the concept of a person in moral theories. I assumed that "person" is always, in one way or another, a relevant concept for any moral theory for the reason that these theories concern the human institution of morality. Ethical theories were taken to be attempts to express what human life involves from an ethical perspective, or what being a moral person is. Moreover, it was assumed that we can discern two aspects in moral theories, although one cannot distinctly separate them from each other. Thus, moral theories are theoretical models for explaining the institution of morality, but they also include a normative aspect in the sense that the theoretical model for understanding the nature of morality is seen as an aid to practical moral reasoning. What connects these two aspects is the conception of the morally relevant. The theoretical model concerning the institution of morality contains a description of the "heart of the matter", that is, of the features of human life which are relevant when they are regarded from the moral point of view. But the concept of the morally relevant was also seen to have a normative bearing for a moral theory in its role as an aid for directing people's moral decision-making.

It was further assumed that the concept of a person is relevant for both of these points of view: the theoretically defined morally relevant was believed to be connected with a description of moral personhood, and the normative aspect of a moral theory was assumed to contain an explication of how one should deliberate and act as a moral person. These assumptions were acceptable if it could be shown that the concepts of a person explicable in various moral theories correspond with the manifest differences between different kinds of moral theories. Thus, a variety of utilitarian moral theories could be said to employ a similar

concept of a person which would deviate from a concept common to different versions of contractarian theories.

These assumptions then served as a basis for a threefold task to explore, first, how the theoretical and the normative conceptions of various moral theories imply the concept of a person; second, which kinds of concepts of a person different moral theories include, and finally, what significance the concept has for understanding the nature of moral theories.

Before starting to analyze particular moral theories it was necessary to make some qualifications concerning the concept of a moral theory. This was made by defining the minimum or necessary conditions any moral theory must fulfil to qualify as such. This minimum condition was found in human intentionality, which was, however, not a sufficient condition for something belonging to the field of morality. Further qualifications were then made by using the classical model for intentional action — the practical syllogism — as an aid. This gave us three possibilities concerning the locus of morality in intentional action: we can place morality in the person either as a moral subject or as a moral object; morality can be seen as a quality of action either through the intended effects of the agent or the realized external effects of the action; or thirdly, the locus of morality can be in the context of intentional action. The context was understood both as the particular situation of the agent and as a wider social context.

The study concentrated on three groups of moral theories classified as utilitarian, contractarian and virtue theories, the contractarian set including both social contract models and a deontological moral theory. The analysis of the three groups of moral theories showed that the initial assumption concerning the importance of the concept of a person in any moral theory was correct. "Person" is a pivotal notion both as far as the theoretical and the normative aspect of any moral theory is concerned. The analysis showed that the theoretically defined morally relevant determines the characteristics to which one must pay attention in one's moral reasoning. In this sense, the theoretical description of the ethically significant was noted to have normative importance in every theory. Further-

more, the study could also establish a direct link between the definition of the morally relevant and the concept of a moral person: these two square with each other in every moral theory under study, irrespective of the theoretical or the normative form of the theory. This observation means that if any two theories represent a similar way of defining the morally relevant, the concepts of a person that they include are similar too. The analysis affirmed that this was the case: the different moral theories in each of the three groups selected for the study represented three different concepts of a person. In each group the theories shared a basically similar notion, whereas this notion was fundamentally different from the concept of any other theory which had been classified in a different group. This result vindicated the construction of three versions of the concept of a moral person: a utilitarian, a contractarian and a narrative concept.

In the following, I will briefly reiterate the main features of these three conceptions by using the two meanings of the Latin word *persona* and the Finnish word for a human face, *kasvot,* as a model of clarification. I will then briefly compare the three models with each other, taking up some points of difference between them. Finally, I will return to the question framed in the introduction to this work concerning the nature of moral theories seen through the concept of a person.

The English word "person" derives from the Latin theatre term *persona*, which refers to the mask an actor uses during a performance. The same word, however, also denotes the one *behind* the mask. I will use this twofold meaning of the term *persona* for discussing the utilitarian and the contractarian concepts of a person. Similarly, I will refer to the meaning of the Finnish word *kasvot* as a useful model for explicating what is involved in the narrative concept of a person.

As we have seen, the utilitarian concept of a person is based essentially on a person's desires. From the utilitarian point of view, desires form a person's real "self" both as a moral and as a non-moral concept. By using the Latin *persona* as a model, the utilitarian conception can be said to represent the real person who is behind any mask of overt behaviour. The utilitarian model defines the person as establishing her person-

hood on certain features of a most fundamental reality. The mask is here formed by whatever people consider themselves to be apart from their desires and preferences which form their real self. The task of a utilitarian moral theory is to explain theoretically what lies behind this mask, that is, what constitutes one's real moral personhood. One can adopt different masks for oneself in the sense that one can attribute one's reasons for action to different sources. But what really is at issue — and this is what the utilitarian theoretical model attempts to show — is that the source of one's actions, whether moral or non-moral, is the set of one's present desires. When one has understood this one can give up all such masks and recognize oneself as a desiring agent, as the kind of a person the utilitarian definition affirms. The utilitarian model is what fundamentally defines, not only what one is, but who one is.

That there is only one true self, the *persona*, squares with the utilitarian concept of the good as primarily non-moral. The moral good is the qualified natural good, and the moral person is the qualified natural person. The perspective of a natural person is that of a "consumer" who is motivated by desire satisfaction. The moral person, again, is a "producer" who sees as her aim the production of the good, the maximization of the utility determined not only by her own desire-satisfaction but by the totality of such utility.

Persona also means the mask itself, and as such it displays the contractarian concept of a person. As a mask the *persona* stands for something given. According to the contractarian moral theories, certain universal features deriving from the necessary conditions of human agency establish moral personhood. Here we can distinguish a different picture: we can see a natural non-moral person who is covered by the mask of moral personhood. The mask of moral personhood conceals one's own individual features as something unimportant. The mask does not destroy these features, it just covers them for the time one acts as a moral person. The masks people wear as moral persons are essentially similar, since the masks represent what is universal. Moral actors do not distinguish each other's natural identities. Such identification is not needed, either; the idea behind being a moral person is just that one hides one's individual

self behind the mask of morality and regards others as wearing one, too. In spite of representing the universal the mask of moral personhood does not stand for what all people have in common. It is an idealized cover explicating the necessary conditions of intentional rational agency. As a natural (at least minimally rational) person everyone understands that these are just the conditions one must accept in any situation that involves other people, that is, in any morally relevant situation, whatever one aims at as a natural person. Moral personhood is, thus, something a natural person accepts as a mask under which she acts according to the role this mask presupposes.

Viewing the moral person as the mask and as the one behind the mask by way of explicating the contractarian and the utilitarian models of moral personhood helps to clarify why the question of personal identity is a central one in utilitarian moral theories whereas the same question receives no attention in the contractarian theories. In the contractarian theories, the given and fixed mask of moral personhood does not allow any changes which would actualize the question of personal identity. Irrespective of people's particular characteristics and their possible temporal changes, their moral personhood remains stable. The situation is different in the utilitarian theories. One's moral personhood is tied to one's desires which admittedly change with time. As one's identity as a (moral) person is determined by such changing desires, it is only natural that the constancy and continuity of one's identity becomes problematic.

Examined against the background of its etymology, the Finnish word for a human face, *kasvot* represents a way of understanding personhood which is radically different from those of Indo-European languages. *Kasvot* is a plural noun deriving from the intransitive verb *kasvaa* (to grow, in its intransitive meaning. *Kasvot* refer to the person in totality, what one has grown to be. There is no mask: *kasvot* cannot be separated from anything. There is no one who emerges from behind a mask: *kasvot* display persons in their very being. The plural form of the word indicates a process or a story: the people are the narrative of their lives, as told by the visible, corporeal form in the *kasvot*. Thus, *kasvot* can be used to exhibit

the idea of the narrative concept of a person. As we have seen, the virtue ethicists' conception of a person involves understanding human life in the form of a story; a story of an individual person embedded in the story of a larger community. The story provides both the aspects of unity and continuity of the person: it links together the past, present, and the future of a person's life. It also gives coherence to features which may otherwise seem mutually incompatible. The form of a story allows great changes to take place both in the psychological structure of a person and in her outer circumstances without endangering her identity.

The form of a story also provides a structure and a meaning for understanding what living as a person involves. It sets a perspective on reality which is not neutral but receives structure from the basic commitments of the person, what she values as good and meaningful. The narrative concept of a person gets a specific colour from the central role the particular is given in virtue theories. The narrative conception of a person does not try to vindicate the particular, but it regards as a universal feature of our personhood that fact that we as persons live particular lives.

The narrative conception of a person can be said to represent a third way of understanding moral personhood, distinct from both the utilitarian and the contractarian ways. The narrative conception widens the perspective from morality to the totality of human life. It tries to encompass everything in human life by allowing a plurality of aspects. As a result, the narrative concept of a person widens into an all-encompassing conception of human life outside of which there can be nothing.

During the study we have been able to distinguish three different conceptions of a person corresponding to three different versions of normative morality and three different approaches to defining the morally relevant. What does this show? Can we use the result for answering the question concerning the nature of moral theories? First, we are justified in saying that since different ethical theories present conceptions of a person fundamentally different from each other, a person is a central theoretical moral notion, whether it is explicated in a theory or not. But, we must ask, do the concepts of a person deviate from each other

because the theories differ from each other? Or is it rather the other way round and the difference between various moral theories derives from the different concepts of a person they include? The correct answer is likely to be that the two, a moral theory and a concept of a person, are inextricably bound up with each other. Moral theories are always formulated for people and of people who regard themselves and who are regarded as moral agents. There cannot be a moral theory which does not include a concept of a person.

On the basis of this study, it is legitimate to suggest that "person" is a most fundamental moral concept in the sense that all moral theories can be conceived as theories about being a person in the moral realm. As such these theories speak to persons in two roles: first, as persons to whom it is shown or proved what the nature of morality is; second, the theories speak to persons recommending or proscribing for them a way for realizing their moral personhood. A moral theory always directs its explication of the institution of morality to a certain kind of a person. This means that "person" is not a theoretically "innocent" concept but that it includes an implicit normative aspect in the sense that it states what is morally important and relevant and what can be left to one side. This brings us to the conclusion that "person" as a moral concept is never a neutral notion.

If "person" is not a normatively neutral notion, we cannot refer to it as a possible solution in our ethical disputes. "Person" as a moral term implies our central normative commitments, it does not offer a neutral ground for solving moral disagreements. The nature of "person" also shows that it is not a univocal concept, at least not in any moral context. We cannot reduce the meaning of the concept to an exhaustive definition everyone should have to accept. Rather, "moral person" can only be used equivocally in the sense that its meaning derives from its context. Consequently, seeking, for example, sufficient or necessary criteria for moral personhood does not help to uncover the essential constituents of personhood: the criteria accepted tend to reflect what is already regarded as crucial for being a moral person. Our conclusion brings us to a further question, beyond the scope of this study: does the concept of a person

used in other than moral contexts have these characteristics? Is "person" always an equivocal, implicitly normative concept? Can we, as persons, ever speak about being a person without attaching some evaluative aspect to our speech?

A last remark concerning the concept of a person and the nature of moral theories is in place. We have noticed that "person" is a very central ethical concept and that we cannot use this concept in a neutral, non-normative way. This conclusion is, actually, a version of Hume's guillotine: one cannot infer values from facts. The concept of a person, a necessary part of every moral theory, is an implicit source of normativeness: "moral person" is never a purely theoretical concept but it always presents a norm for those who care to listen to what the moral philosopher wishes to say.

Bibliography

Adams, E. M.

1984 "The subjective normative structure of agency." *Gewirth's Ethical Rationalism: Critical Essays with a Reply by Alan Gewirth*, 8–22. Ed. by Edward Regis Jr., The University of Chicago Press, Chicago, London.

Almond, Brenda

1990 "Alasdair MacIntyre: the virtue of tradition." *Journal of Applied Philosophy.* 99–103.

Anscombe, G. E. M.

1958 "Modern moral philosophy." *Philosophy,* 1–19.

Aristotle

1953 *Metaphysics. A Revised Text with Introduction and Commentary I-II.* W. D. Ross, Oxford University Press, Oxford.

1990 *Nichomachean Ethics.* Translated by H. Rackham, Harward University Press, Cambridge, Mass., London.

Brandt, Richard

1959 *Ethical Theory.* Prentice Hall, Englewood Cliffs, New Jersey.

1979 *A Theory of the Good and the Right.* Clarendon Press, Oxford.

1988 "The structure of virtue." *Ethical Theory: Character and Virtue.* Midwest Studies in Philosophy. Volume XIII, 64–82. University of Notre Dame Press, Notre Dame.

Carruthers, Peter

1986 *Introducing Persons: Theories and Arguments in the Philosophy of Mind.* Croom Helm, London & Sydney.

Churchland, Patricia Smith

1989 *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* MIT Press, Cambridge, Mass.

Clowney, David

1990 "Virtues, rules, and the foundations of ethics." *Canadian Journal of Philosophy*, 49–68.

Cockburn, David

1987 "Critical notice." *Philosophical Investigations,* 54–72.

DAVIDSON, DONALD
1980 *Essays on Actions and Events*. Clarendon Press, Oxford.

FOOT, PHILIPPA
1978 *Virtues and Vices and Other Essays in Moral Philosophy*. Basil Blackwell, Oxford.

FRANKENA, WILLIAM K.
1963 *Ethics*. Englewood Cliffs, N. J., Prentice-Hall.
1973 *Ethics*. 2nd ed. Englewood Cliffs, N. J., Prentice-Hall.
1983 "MacIntyre and modern morality." *Ethics*, 579–587.

GAUTHIER, DAVID
1988 *Morals by Agreement*. Clarendon Press, Oxford.

GEACH, PETER
1977 *The Virtues*. Cambridge University Press, Cambridge, London, New York, Melbourne.

GEWIRTH, ALAN
1978 *Reason and Morality*. University of Chicago Press, Chicago.

GILLET, GRANT
1986 "Brain bisection and personal identity." *Mind*, 224–229.
1987 "Reasoning about persons." *Persons and Personalities*, 75–88. Ed. by Arthur Peacocke & Grant Gillet. Basil Blackwell, Oxford.

GLOVER, JONATHAN
1988 *I: The Philosophy and Psychology of Personal Identity*. Penguin Books, Harmondsworth.

GRIFFIN, JAMES
1990 *Well-being: Its Meaning, Measurement and Moral Importance*. Clarendon Press, Oxford.

HARE, RICHARD M.
1952 *The Language of Morals*. Oxford University Press, Oxford.
1963 *Freedom and Reason*. Oxford University Press, Oxford.
1984 *Moral Thinking: Its Levels, Method and Point*. Clarendon Press, Oxford.

HARSANYI, J.C.
1988 "Problems with act-utilitarism." *Hare and Critics: Essays on Moral Thinking*, 89–99. Ed. by Douglas Seanor & N. Fotion, Clarendon Press, Oxford.

HUDSON, STEPHEN D.
1986 *Human Character and Morality*. Routledge & Kegan Paul, Boston.

HUMAN BEINGS
1991 *Human Beings*. Ed. by David Cockburn, Royal Institute of Philosophy

Supplement: 29, Cambridge University Press, Cambridge.

HUME, DAVID

1984 *A Treatise of Human Nature.* Edited with an introduction by Ernest C. Mossner, Penguin Books, Harmondsworth.

1990 *Enquiries Concerning Human Understanding and Concerning the Principles of Morals.* Reprinted from the 1777 edition with Introduction and Analytical Index by L. A. Selby-Bigge, 3rd ed., with text revised and notes by P. H. Nidditch, Clarendon Press, Oxford.

KENNY, ANTHONY

1978 *Aristotelian Ethics.* Oxford University Press, Oxford.

KUHN, THOMAS

1962 *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago.

KUKATHAS, CHANDRAN & PETTIT, PHILIP

1990 *Rawls: A Theory of Justice and Its Critics.* Polity Press, Cambridge.

LOCKE, DON

1987 "The Parfit population problem." *Philosophy*, 131–157.

LOUDEN, ROBERT B.

1984 "On some vices of virtue ethics." *American Philosophical Quarterly*, 227–235.

1990 "Virtue ethics and anti-theory." *Philosophical Studies*, 93–114.

MACINTYRE, ALASDAIR

1987 *After Virtue: A Study in Moral Theory.* 2nd ed., Duckworth, London.

1988 *Whose Justice? Which Rationality?* Duckworth, London.

1990 *Three Rival Versions of Moral Enquiry.* Duckworth, London.

MACKIE, JOHN

1990 *Ethics: Inventing Right and Wrong.* Penguin Books, Harmondsworth.

MADELL, GEOFFREY

1981 *The Identity of the Self.* Edinburgh University Press, Edinburgh.

MARKHAM, IAN

1991 "Faith and reason: reflections on MacIntyre's 'tradition-constituted enquiry'." *Religious Studies,* 259–267.

MCNAUGHTON, DAVID

1988 *Moral Vision: An Introduction to Ethics.* Basil Blackwell, Oxford and New York.

MCSHEA, ROBERT J.

1990 *Morality and Human Nature: A New Route to Ethical Theory.* Temple University Press, Philadelphia.

MURDOCH, IRIS
1967 *The Sovereignty of Good over Other Concepts.* The Leslie Stephen Lecture 1967. Cambridge University Press, Cambridge.

NAGEL, T.
1988 "Foudations of impartiality." *Hare and Critics: Essays on Moral Thinking,* 101–112. Ed. by Douglas Seanor & N. Fotion, Clarendon Press, Oxford.

NIELSEN, KAI
1984 "Against ethical rationalism." *Gewirth's Ethical Rationalism. Critical Essays with a Reply by Alan Gewirth,* 59–83. Ed. by Edward Regis Jr., The University of Chicago Press, Chicago, London.

NUSSBAUM, MARTHA C.
1986 *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy.* Cambridge University Press, Cambridge.
1990 *Love's Knowledge: Essays in Philosophy and Literature.* Oxford University Press, New York and Oxford.
1992 "Virtue revived." *Times Literary Supplement*, July 3, 9–11.

OAKLANDER, L. NATHAN
1987 "Parfit, circularity, and the unity of consciousness." *Mind*, 525–529.

PARFIT, DEREK
1984 *Reasons and Persons.* Clarendon Press, Oxford.

PENCE, GREGORY E.
1984 "Recent work on virtues". *American Philosophical Quarterly*, 281–297.

POGGE, THOMAS W.
1989 *Realizing Rawls.* Cornell University Press, Ithaca, London.

PUTNAM, HILARY
1983 "There is at least one *a priori* truth." *Realism and Reason: Philosophical Papers, vol. III*, 98–114, Cambridge University Press, Cambridge.

PUCCETTI, ROLAND
1980 "The duplication argument defeated." *Mind*, 582–586.

RAWLS, JOHN
1972 *A Theory of Justice.* Clarendon Press, Oxford.
1993 *Political Liberalism.* Columbia University Press, New York.

RICHARDS, DAVID A.
1988 "Prescriptivism, constructivism and rights." *Hare and Critics: Essays on Moral Thinking,* 101–112. Ed. by Douglas Seanor & N. Fotion, Clarendon Press, Oxford.

RICŒUR, PAUL
1990 *Soi-même comme un autre.* Seuil, Paris.

RORTY, AMELIE OKSENBERG
1988 *Mind in Action: Essays in the Philosophy of Mind.* Beacon Press, Boston.

SANDEL, MICHAEL
1982 *Liberalism and the Limits of Justice.* Cambridge University Press, Cambridge.

SCANLON, T.M.
1988 "Levels of moral thinking." *Hare and Critics: Essays on Moral Thinking,* 129–146. Ed. by Douglas Seanor & N. Fotion. Clarendon Press, Oxford.

SEANOR, DOUGLAS & FOTION, N.
1988 "The levels, methods and points." *Hare and Critics: Essays on Moral Thinking,* 3–8. Ed. by Douglas Seanor & N. Fotion. Clarendon Press, Oxford.

SHERMAN, NANCY
1991 *The Fabric of Character: Aristotle's Theory of Virtue.* Clarendon Press, Oxford.

SHOEMAKER, SYDNEY
1985 "Critical notice." *Mind*, 443–453.

SHOEMAKER, SYDNEY & SWINBURNE, RICHARD
1984 *Personal Identity.* Blackwell, Oxford.

SINGER, PETER
1988 "Reasoning towards utilitarism." *Hare and Critics: Essays on Moral Thinking,* 147–159. Ed. by Douglas Seanor & N. Fotion. Clarendon Press, Oxford.

SPERRY, R. W.
1966 "Brain bisection and the mechanisms of consciousness." *The Brain and Conscious Experience.* Ed. by J. C. Eccles, Springer-Verlag, Berlin & New York.

SPRIGGE, T. L. S.
1988 "Personal and impersonal identity", *Mind*, 29–49.

STOCKER, MICHAEL
1990 *Plural and Conflicting Values.* Clarendon Press, Oxford.

STROLL, AVRUM
1972 "Identity." *The Encyclopedia of Philosophy,* 121–124. Collier MacMillan Publishers, New York, London.

TAYLOR, CHARLES

1989 *Sources of the Self: Making of the Modern Identity*. Cambridge University Press, Cambridge.

1990a *Human Agency and Language: Philosophical Papers 1*. Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney.

1990b *Philosophy and the Human Sciences: Philosophical Papers 2*. Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney.

TAYLOR, C. C. W.

1987 "Critical notice. *The Fragility of Goodness. Luck and Ethics in Greek Tragedy and Philosophy*. By Martha C. Nussbaum." *Mind,* 407–414.

THEORIES OF ETHICS

1967 *Theories of Ethics*. Ed. by Philippa Foot, Oxford University Press, Oxford.

TRIANOSKY, GREGORY

1990 "What is virtue ethics all about?" *American Philosophical Quarterly*, 335–344.

TRUPP, ANDREAS

1987 *Why We Are not What We Think We Are: A New Approach to the Nature of Personal Identity and of Time*. Verlag Peter Lang, Frankfurt am Main, Bern, New York, Paris.

WALLACE, JAMES D.

1978 *Virtues and Vices*. Cornell University Press, Ithaca and London.

1988 "Ethics and the craft analogy." *Ethical Theory: Character and Virtue*. Midwest Studies in Philosophy, XIII, 222–232. University of Notre Dame Press, Notre Dame.

1991 "Theorizing about morals". Noûς, 176–187.

WARNOCK, G. J.

1981 *Contemporary Moral Philosophy*. MacMillan, London.

WESTON, M.

1991 "Three rival versions of moral enquiry." *Mind*, 400–403.

WIGGINS, DAVID

1967 *Identity and Spatio-Temporal Continuity*. Basil Blackwell, Oxford.

1987 *Needs, Values, Truth: Essays in the Philosophy of Value*. Basil Blackwell, Oxford.

WILLIAMS, BERNARD

1973 *Problems of the Self*. Cambridge University Press, Cambridge.

1987 *Ethics and the Limits of Philosophy*. Fontain Press, London.

WOLF, URSULA
1986 "Was es heißt, sein Leben zu leben." *Philosophische Rundschau,* 242–265.
WRIGHT, G. H. VON
1968 *The Varieties of Goodness.* Routledge & Kegan Paul, London.

Index of names