Utilitarianism: Problems, Recent Solutions, Possible Developments

Luca Stroppa

857116

Supervisor: Prof. Renato Pettoello

2016/2017

To Maurizio Fiorani

Acknowledgments

I discovered Derek Parfit, and consequently the majority of the bibliography of this work, thanks to Jacopo Freri. He introduced me to the problems of Personal Identity, and many of my ideas about Population Ethics come from discussions with him. I owe much to him.

I also owe a great deal to Giuliana Mancuso. She followed my bizarre interests for more than a year, with praiseworthy effort and the greatest dedication. She showed patience even when facing the most obscure pages I wrote, gave me important suggestions, not only about this work, and taught me the importance of joining precision with clearness when exposing an argument.

Guido Tonella, Federico Faroldi and Tim Campbell read and objected what has become Part 3 of this work, which is the part less historical and more speculative. They all suggested me books and papers with which I developed my researches, and they prompted my interests with discussions. Together with Jacopo Freri and Giuliana Mancuso, they influenced my work and my enthusiasm towards it much more than what it is possible to report in the acknowledgments.

I have to thank Prof. Renato Pettoello for accepting my thesis, for suggesting me Giuliana Mancuso's help and for his remarkable willingness to discuss about every subject of philosophy with his students.

During the writing of this work I was supported, in a less academic manner, also by members of my family and by friends not mentioned previously. They are too many to be listed here. I thank them all, and I list here only the ones that, for different reasons, had evident impact on this work: my parents Enrica and Giuseppe, my siblings Alice and Francesco, my Grandma Elena, Alessandro Torchio, Arianna Bettin Campanini, Beatrice Foti, Clara Fenocchi, Daniele Lepore, Giovanni Colpani, Luca Savioli and, last only alphabetically, Michele Piccoli.

I also have to thank my professors at the Liceo, Isa Veluti and Stefania Landi. Without them, I am not sure I would have studied philosophy.

Index

Preface

"At last the horizon appears free to us again, even granted that it is not bright; at last our ships may venture out again, venture out to face any danger; all the daring of the lover of knowledge is permitted again; the sea, *our sea*, lies open again; perhaps there has never been such an 'open sea'."

The quote above is what Nietzsche wrote in the final part of the § 343 from the fifth book of *The Gay Science*. The title of the aphorism is "on what Cheerfulness signifies". It is about the death of God, and the collapse of everything related to it, "for example", as there stated by Nietzsche, "European morality". In the aphorism, philosophers and "free spirits" cheer because, free from God's oppression, they can dare to think against His precepts, in order to rebuilt what has been destroyed in God's fall. This means, for example, that European morality has to be rethought, and Nietzsche's proposal is a new kind of morality based on the prevalence of the stronger on the weaker.

Apparently, no thinker seems more unrelated to the aphorism above than Derek Parfit. Known for his attention for weaker people, remarkable for his faith in human reasoning and for his belief in the possibility to establish an indubitable and universally shared morality, Parfit seems not linked in any way to *The Gay Science*'s aphorism § 343.

But the very first page of his first book is entirely occupied by the quote above. Furthermore, by a strange coincidence that clearly Parfit could not foresee, also in the very last page of the last book he published before recently dying appears the same Nietzsche's quote. Somehow, Parfit felt a strong connection between himself and the aphorism § 343, and Destiny, if there is such entity, highlighted this connection. What does this connection consists of? Are Parfit's and Nietzsche's horizons and seas the same horizon and the same sea? What are the similarities, and what are the differences between their widely free views?

In order to answer this question we must travel a long path, and make surprising discoveries. Among them, we will see the solution of what has been considered the profoundest problem in ethics, our most common intuitions regarding ourselves will be deeply shaken and the apparently indubitable view according to

which an action is bad if it affects people for the worse will be recognized as unfit for most part of moral thinking.

This path begins in the Victorian age with Henry Sidgwick, passes through Derek Parfit and has an end in a future not foreseeable yet. Once analyzed the problems that Sidgwick poses, Parfit's answers and some possible developments of their thesis, we will understand the relationship between Parfit and Nietzsche's aphorism § 343.

Introduction

Born in 1942 in Chengdu, China, but grown in Oxford, Derek Parfit became famous in 1971 with a paper on *Personal Identity*,[1] and his fame has grown ever since. He graduated in history at Balliol College, Oxford, in 1964, and three years later he gained a prize fellowship to All Souls College, in Oxford, where he switched studies from history to philosophy and wrote the mentioned paper, several articles and books. In his lifetime he received several awards for his philosophical works, among which the Rolf Schock Prizes in 2014 for Logic and Philosophy. [2] He died the first of January 2017.[3] He is currently regarded as one of the major moral philosophers of the last hundred years.[4]

In his life, he wrote only two books, and both led to a great discussion among contemporary philosophers. The most recent of them is *On What Matters*, in three volumes. The first two volumes were published in 2011, whereas the third has been published posthumous in March 2017. In the last pages of Volume Three Parfit wrote he hoped to write a fourth volume, which he did not for clear reasons. *On What Matters* discusses several ethics theories, such as the Kantian theory, the theories of contractarianism - particularly Scanlon's one - , and the utilitarian theory, which I will explain in detail later; Parfit tried to show how those theories converge and how they seem likely to become a single theory; also, he discussed other questions of meta-ethics such as the relationship between normativity and motivation in ethics.

His other book is *Reasons and Persons*, first published in 1984 and reprinted with corrections in 1986 and 1987. One of the aims of the present work is to explain

[1] The article is (Parfit, Personal Identity, 1971)
[2] See (Royal Swedish Academy of Science, 2014)
[3] I do not find a deeper analysis of Parfit's life particularly useful for this work, since none of its events seems to have a relevant relation with his philosophical views. I also find that Parfit's theory on Personal Identity and several of his declarations, for example the beginning of (obituary, 2017), seem to discourage any analysis of this kind. According to (MacFarquhar, 2011), Parfit himself had difficulty in remembering his own past and rarely thought about it: this detail of his personality dissuades me from thinking that his biography has any meaning for this work. Anyhow, for a full report of his biography I recommend (Grimes, 2017). For a more analytic, but less complete article, which is less complete only because written before the death of the philosopher, I suggest (MacFarquhar, 2011) .
[4] At least according to (O'Grady, 2017) , (Matthews, 2017) , several articles quoted in (Justin, 2017) and countless others.

and discuss the main concepts of this book, if not in their full depth, at least in a clear manner.

The book is divided in four parts. The first part, entitled "Self-Defeating Theories", focuses on reason applied to ethics, or, in other words, on establishing criteria for a coherent and effective method for ethical reasoning. The second part, entitled "Rationality and Time", underlines the incoherence of a theory that has its paramount aim in self-interest. Here Parfit tries to show this theory's irrationality by challenging it through its confrontation with an altruist theory and an artificial theory that has his paramount aim in the immediate satisfaction of present aims. In this second part Parfit writes that, without a new concept of personal identity, no confutation of a self-interested theory can be theoretically satisfying. The third part, entitled "Personal Identity", states this new concept, according to which personal identity does not matter when we act: in this way, the self-interest theory can be rejected. This new concept of personal identity is in clear conflict with our most common assumptions about ourselves. The fourth and last part of *Reasons and Persons*, whose title is "Future Generations", adds new reflections on personal identity and inspects our responsibility towards others, particularly future generations. The results are possibly even more surprising than the ones on personal identity: Parfit himself does not manage to solve the moral problems raised in this part of his book, because every solution seems to lead to a paradox, or to conclusions that Parfit finds repugnant. This part of *Reasons and Persons* rises so many question that an entire new field of moral philosophy, called Population Ethics, arose in order to solve those problems.

Reasons and Persons was conceived as a try to solve problems arisen in one of the most important books about morality of the Victorian age. Therefore, in order to correctly understand *Reasons and Persons* I need to explain satisfactorily some concepts from this book, that is *The Methods of Ethics*, by Henry Sidgwick (1838-1900).

Parfit himself seems to encourage the reading of Sidgwick's *Methods of Ethics* for comprehending his book, since he explicitly considered Henry Sidgwick to

be his "master".[5] Sidgwick's *Methods* is considered one of the most clear and complete exposition of utilitarian thinking,[6] and Parfit believed it to be the best book on Ethics ever written.[7] In this book Sidgwick examines the main methods of thinking ethical behavior. He analyzes them in their distinguishing features, in their convergences and their divergences. For our purpose, a very important result of this work will be the unavoidable contrast between self-interest and altruism. How to avoid this contrast is considered by Sidgwick "the profoundest problem in Ethics".[8]

Parfit's *Reasons and Persons* aims, in effects, to solve Sidgwick's "profoundest problem". But Parfit is not the first to do so: Thomas Nagel's *The Possibility of Altruism* can be considered the most important book on this matter before Parfit's one, and had some influence on this latter. Nagel examines why people are motivated to perform certain acts that will benefit them in the future, and states that an important part of the answer can be found in the concept of personal identity, conceived as constant in time. After that he tries to find how people might be motivated to benefit not only themselves, but also others. He is not able to resolve the contrast between self-interest and altruism. He concludes that, in order to give a convincing answer to this question, we need another concept of personal identity. This concept will be found by Parfit in the third part of *Reasons and Persons*.

---

[5] In (Parfit, 2011 a, p. xl) Parfit describes his debts to Sidgwick as follows: "Of my reasons for becoming a graduate student in philosophy, one was the fact that, in wondering how to spend my life, I found it hard to decide what really matters. I knew that philosophers tried to answer this question, and to become wise. It was disappointing to find that most of the philosophers who taught me, or whom I was told to read, believed that the question 'What matters?' couldn't have a true answer, or didn't even make sense. But I bought a second-hand copy of Sidgwick's book, and I found that he at least believed that some things matter. And it was from Sidgwick that I learnt most about the other questions that moral philosophers should ask, and about some of the answers." In (Parfit, 2011 a, p. xxxiii) Sidgwick is presented as Parfit's "master" together with Kant. But Kant can be considered Parfit's "master" only in the three *On What Matters* volumes: Parfit started to appreciate and study Kant's work, particularly the *Grundlegung zur Metaphysik der Sitten*, only after 1990. Before that, Parfit was "deeply opposed both to some of Kant's main claims, and to his way of doing philosophy" and nearly ignored him, as stated in (Parfit, 2011 a, p. xli), where he also recommend the reader to do unlike him and read some of Kant's books. But it must be noticed that, if we accept Nagel's analysis of Kant in (Nagel, 1978, p. 11-12), Parfit seems to share, in (Parfit, 1984), if not Kant's ideas, at least his desire to "discover requirements on actions which apply to a man on no condition about what he wants, how he feels etc." at the moment of the action. I will better explain Nagel's view on Kant later, at note 238.

[6] For example, see (Rawls, 1981). To be fair, as we will soon see, (Sidgwick, 1874) aims *not* to be a book on Utilitarianism, neither its final thesis can be described satisfactorily as purely Utilitarian.

[7] (Parfit, 2011 a, p. xxxiii)

[8] (Sidgwick, 1874, p. 386 note 4)

The three books, particularly Sidgwick's and Parfit's book, will be commented and explained, but they will be not analyzed in their completeness, because that would require too much space and would not help much four our purposes. Thus large parts of those books of great importance for modern moral philosophy, such as *The Method'*s chapter on free will,[9] will be ignored.

This work is divided in three parts: "Problems", "Recent Solutions" and "Possible developments". In the first part, that is an analysis of *The Methods of Ethics*, some concepts of Sidgwick's theory of morality will be explained, and the most important theoretical problems of this theory will be stated. One of those problems is the contrast between altruism and self-interest, that Sidgwick considered "the profoundest problem in ethics", which will be the core of the second part of this work. Another problem is the first problem of Population Ethics raised by a philosopher, and it will be discussed in the third part of this work. The second part consists in the analysis of the solutions found to the problem of the contrast between altruism an self-interest. Since those solutions have less than 50 years, this part is called "Recent Solutions". It consists in a quick analysis of Nagels' *The Possibility of Altruism*, the first book which tries to cope with "the profoundest problem in Ethics", and an analysis of Parfit's *Reasons and Persons*. The third part is a development of Parfit's theories about Personal Identity and responsibility towards future generations. This section is called "Possible Developments" and consists in a new proposal of some thesis in Population Ethics that have Parfit's work as a basis. In what follows I explain more in detail the contents of this work.

Part 1, "Problems", includes four chapters. The first analyzes Sidgwick's conception of *The Methods of Ethics*, his aim of solving the contrast between self-interest and altruism and the core concepts of Sidgwick's book. In chapter 2 it will be given a report of Sidgwick's analysis of the theory according to which it is supremely rational for an agent to benefit herself, a theory called Egoism, and it will be showed how Egoism is incompatible with any kind of Altruism. The theory according to which the agent ought to conform to certain rules unconditionally prescribed, called Intuitionism, and the limits of this theory as highlighted by

[9] That can be found in (Sidgwick, 1874, p. 57-76)

Sidgwick, will be the subject of chapter 3. Chapter 4 reports Sidgwick's analysis of the theory according to which the agent ought to maximize pleasure over pain, called Utilitarianism, and of its limits. It will be showed how those limits are overcome if Utilitarianism is unified with Intuitionism, and how them both are incompatible with Egoism. It will be pointed out how Sidgwick's union of Intuitionism and Utilitarianism can be identified with Consequentialism. At the end of chapter 4 the problems of Utilitarianism and Consequentialism highlighted in this Part 1 will be listed. Two of those problem are the incompatibility between Egoism and Altruism and the fact that it is not clear whether, in maximizing pleasure over pain, a Utilitarian ought to aim for the highest total pleasure or the highest average pleasure.

Part 2, "Recent solutions", includes three chapters and is devoted to reporting how Nagel and Parfit tried to solve the incompatibility between Egoism and Altruism. In chapter 5 it is pointed out how Nagel demonstrates, in *The Possibility of Altruism*, that reasons for acting are intertemporal, since the agent conceives herself as a being constant in time. Unfortunately, Nagel is not able to demonstrate that the agent should conceive herself as someone among others: if this is demonstrated, reasons for acting would be interpersonal, and the incompatibility between Egoism and Altruism would be overcome. Nagel guesses that, for such a demonstration, a new theory of Personal Identity is required. In chapter 6 it is illustrated how in *Reasons and Persons* Parfit challenges Egoism with objections that come from Consequentialism and from a fictional theory according to which the agent ought to satisfy only her present aims. In chapter 7 it is examined Parfit's argument for stating that Personal Identity consists in psychological connectedness, which is the holding of the same psychological features for a certain amount of time, and psychological continuity, which is the holding of overlapping chains of connectedness. Furthermore, Parfit points out how only connectedness matters to the agent, and not identity: since connectedness changes in an agent's life and the agent's identity does not matter, it is not supremely rational for an agent to benefit herself. Therefore Egoism should be rejected.

Part 3, "Possible developments", includes three chapters and is the most original part of this work, since here the reconstruction makes way for a new proposal. This part is devoted to solve the problem of the uncertainty of

Utilitarianism between maximizing the highest total pleasure or the highest average pleasure. The beginning of chapter 8 is an introduction to Population Ethics, which is the field of ethics that seeks a theory of population value. It will be pointed out how, if Parfit's theory of Personal Identity is reliable, the conclusion of a theory of Population Ethics must be fit for many individual choices as well, since also in these choices there is uncertainty between maximizing the highest total pleasure or the highest average pleasure. Since it has been showed how some extremely appealing adequacy conditions for a theory of population value are incompatible, and how therefore it might be impossible to find such a theory, in chapter 9 it will be tested if two of these adequacy conditions are fit for individual choices: if one of them is unfit for individual choices, it is probably unfit for choices between populations. In chapter 9 the condition known as *Quality Condition*, according to which "there is at least one perfectly equal population with very high welfare which is at least as good as any population with very low positive welfare", is translated in terms of individual choices and found doubtful, since it seems to rely on the human difficulty of grasping great numbers. If the *Quality Condition* has to be rejected, a Utilitarian ought to maximize the highest total pleasure, and not the highest average pleasure. In chapter 10 some considerations will be made about how a reliable theory of probability might help in Population Ethics, and how these considerations on probability support the claim according to which a Utilitarian ought to maximize the highest total pleasure.

It will follow a Conclusion where the starting hint to the relationship between Parfit and Nietzsche will be considered again and the importance of Population Ethics will be stressed.

Part 1: Problems

Chapter 1: The main concepts of *the Methods of Ethics*

1.1. Sidgwick's Aim

In his foreword to the 1981 edition of *The Methods of Ethics* John Rawls writes that Sidgwick's treatise is important in moral philosophy for three motives. First, it is "the clearest and most accessible formulation of what we may call 'the classical utilitarian doctrine'"; second, "Sidgwick is more aware than other classical authors of the many difficulties this [Utilitarian] doctrine faces, and attempts to deal with them in a consistent and thorough way while never departing from the strict doctrine." Third, "It is the first truly academic work in moral philosophy which undertakes to provide a systematic comparative study of moral conceptions".[10]

Sidgwick himself was not nearly as enthusiastic as Rawls was. About his masterpiece in 1870 he wrote to a friend: "The book solves nothing, but may clear up the ideas of one or two people, a little."[11]

All four sentences, the three by Rawls and the last by Sidgwick, deserve to be commented. A good way for commenting them is by summarizing how Sidgwick thought to write "a systematic comparative study of moral conceptions". We will discover how, even if the first two statements by Rawls describe Sidgwick as a particularly coherent Utilitarian, it should not be overlooked that Sidgwick did not think at himself as a pure Utilitarian. He did not write his book with the purpose of dealing with any difficulty of the Utilitarian theory; in fact, it might be argued that he raises as many objections to Utilitarianism as arguments for its defence. But this last matter will result clear when discussing the final part of the work.

As Sidgwick himself writes in an account on the conception of his treatise, his "first adhesion to a definite Ethical system was to the Utilitarianism of Mill",[12] even if he was "not satisfied with the Utilitarian method of dealing with the conflict

[10] (Rawls, 1981, p. v)
[11] This is reported in (Parfit, 2011 a, p. xxxiv), cf. A. Sidgwick, E. Sidgwick, *Henry Sidgwick. A Memoir*, MacMillan and Co., London 1906, p. 284.
[12] (Sidgwick, Methods of Ethics, 1874, p. xvii). I will explain the term Utilitarianism soon.

between Interest and Duty".[13] In other words, he perceived a "profound discrepancy between the natural end of action - private happiness [or Interest], and the end of duty - general happiness",[14] and did not found Mill's solution of this discrepancy, that relied on heroism,[15] philosophically satisfying. The alternative to Utilitarianism, that is Intuitionism, appeared to Sidgwick too imprecise in definitions and axioms at first. The notion of Intuitionism will be better developed later, but I define it provisionally here as the view according to which at least some of the basic moral propositions are self-evident. Sidgwick's first contact with a book about Intuitionism was William Whewell's *Elements of Morality*.[16] Since Intuitionism seemed imprecise he adhered to Utilitarianism, despite the problem of the discrepancy of interest and duty.

Nevertheless, Sidgwick was not incline to overlook the problem of this discrepancy. In order to overcome it he was "forced to recognize the need of a fundamental ethical intuition"[17] and read Kant's ethic, appreciating "Kant's ethical principle rather than its metaphysical basis [. . .].That whatever is right for me must be right for all persons in similar circumstances - which was the form in which I accepted the Kantian maxim - seemed [. . .] certainly fundamental, certainly true, and not without practical importance."[18]

Sidgwick saw this as a step forward, but not the final step at all. The problem of the discrepancy between Duty and Interest was not solved, "for the Rational Egoist - a man who had learnt from Hobbes that Self-preservation is the first law of Nature and Self-interest the only rational basis of social morality [. . .] might accept the Kantian principle and remain an Egoist."[19]

---

[13] Ibid.
[14] (Sidgwick, 1874, p. xv)
[15] What I mean is that Sidgwick writes "I put aside Mill's phrases that such sacrifice [the sacrifice of the agent's happiness for the sake of someone else's happiness] was "heroic": that it was not "well" with me unless I was in a disposition to make it" (Sidgwick, 1874, p. xvi). Actually Mill did more than just encouraging to perform "heroic" acts: he wrote a theoretical argument to reduce the discrepancy between egoism and altruism, but Sidgwick rejects it. A summary of Mill's argument an Sidgwick's objections can be found in (Sidgwick, 1874, p. 387-388)
[16] As stated in (Sidgwick, 1874, p. xvii)
[17] (Sidgwick, 1874, p. xviii)
[18] (Sidgwick, 1874, p. xix)
[19] Ibid.

At this point Sidgwick read Joseph Butler's *Sermons*[20] and "found in him [. . .] a view very similar to that which had developed itself in my own mind in struggling to assimilate Mill and Kant. [. . .]That is, he recognised a 'Dualism of the Governing Faculty' or as I prefer to say 'Dualism of the Practical Reason'"[21]. This Dualism is between Morality and Egoism, and is considered by Sidgwick, as I previously stated, the profoundest problem of Ethics.

Butler's *Sermons* carried Sidgwick "a further step away from Mill"[22] and made him become "an Intuitionist to a certain extent."[23] He accepted intuitionist's views "For the supreme rule of aiming at the general happiness [. . .] must rest on a fundamental moral intuition".[24] Sidgwick decided then to inspect this moral intuition. He writes: "I had to read Aristotle again; and a light seemed to dawn upon me as to the meaning and drift of his procedure - especially in Books ii., iii., iv. of the *Ethics* [. . .]. What he gave us there was the Common Sense Morality of Greece, reduced to consistency by careful comparison: given not as something external to him but as what 'we' - he and others - think, ascertained by reflection. [. . .] Might I not imitate this: do the same for our morality here and now, in the same manner of impartial reflection on current opinion?"[25] Thus he started to write about it; the result was what then would have become the first eleven chapters of the third book of *The Methods of Ethics*.

I will not examine the process through which, from those eleven chapters, Sidgwick decided to expand his work so to examine not only the method of Common Sense Morality but also the Utilitarian and Egoist methods. What I want to stress is how he was attracted and interested, in different ways and in different times, by all

[20] Joseph Butler (1692-1752) was an author remarkable for his contribution to ethics and theology. His most famous theological book is probably *Analogy of Religion, Natural and Revealed* (1736), whose main contents is an attack on theism, and whose first appendix is an attack on Locke's view on personal identity. Butler's masterpiece in Ethics are considered the *Fifteen Sermons Preached at the Rolls Chapel* (1726). In Ethics, he is particularly important for his critics on Thomas Hobbes' view on the egoist human nature. In fact, according to Butler, selfish feelings are not the only components of human motivation: also benevolence has an important role. In addition, the faculty of conscience, that is an inborn sense of right and wrong given by God, is capable to analyze between the components of human motivation and choose between them. For further information on Butler's ethics see https://plato.stanford.edu/entries/butler-moral/.
[21] (Sidgwick, 1874, p. xx-xxi)
[22] (Sidgwick, 1874, p. xxi)
[23] Ibid.
[24] Ibid.
[25] (Sidgwick, 1874, p. xxi-xxii)

three methods: this is the reason why Sidgwick decided to "provide a systematic comparative study" of those methods.

As we have seen, Rawls defines Sidgwick's book as "the clearest and most accessible formulation of what we may call 'the classical utilitarian doctrine'". Since Sidgwick did not considered himself less intuitionist than utilitarian, and since Intuitionism and Utilitarianism seems incompatible, being the former based on the principles for moral actions and the latter based on the consequences of the moral action, some clarifications are needed in order to correctly comprehend this definition by Rawls. Sidgwick's self-evaluation as a partial intuitionist does not imply that his moral theory diverges in any way from classical Utilitarianism; even less it is implied that his exposition of Utilitarian thinking diverges from classical Utilitarianism. Sidgwick's Utilitarian theory is not undermined by intuitional influences, but rather expanded and completed by them. If this is true for his theory, the same cannot be said, as we will see examining the conclusion of the treatise, of Sidgwick's actual moral practice, in which he adheres, unless exceptional cases, to the Morality of Common Sense rather than to what classical Utilitarianism would recommend. What are those "exceptional cases" will be clearer in 4.2.

Sidgwick's account has given us hints on why, according to Rawls, he was "more aware than other classical authors of the many difficulties this doctrine [which is Utilitarianism] faces": Sidgwick thinks that Utilitarianism is incomplete, and tries to fix its incompleteness through other theories. He does not manage to resolve all the problems of the Utilitarian theory.

In fact, Sidgwick's book solved "nothing", as Sidgwick himself claims, because he did not manage to overcome the hiatus between Interest and Duty, leaving the Dualism of Practical Reason (that is, as the reader will remind, considered by Sidgwick the profoundest problem of ethics) as it were.

The Methods of Ethics is divided in four books. The first does not have any title and contains mainly an examination of some concepts of great interest in moral reasoning. Each of the other books contains an analysis of one of the three above-mentioned moral theories and is titled before the moral theory analyzed: consequently, the second book is entitled *Egoism*, the third *Intuitionism* and the last *Utilitarianism*. This last book contains also a *Concluding Chapter*, in which

Sidgwick considers the relations between the three theories. I will not always follow Sidgwick's order in my exposition.[26]

In the first book of the *Methods,* the conclusions that are more helpful in order to understand Sidgwick's book, the Utilitarian theory and its limits concern the following concepts: *Ethics, Intuitionism, Pleasure, Pain* and *Reason*.

1.2. Sidgwick's *Ethics* and its *Methods*

The first concept analyzed in Sidgwick's book is that of Ethics. He defined a "method of ethics" (henceforth Method) as "any rational procedure by which we determine what individual human beings "ought" or what it is "right" for them to do, or to seek to realise by voluntary action".[27] Thus we might state Sidgwick's conception of Ethics, as expressed at the beginning of the *Methods of Ethics,* as follows:

> *Ethics (according to Sidgwick)*: "the science or study of what is right or what ought to be, so far as this depends upon the voluntary action of individuals".[28]

The term "individual" marks the distinction between ethics and politics, because this latter "seeks to determine the proper constitution and the right public conduct of governed societies".[29] Ethics and politics are "distinguished from positive sciences by having as their special and primary object to determine what ought to be, and not to ascertain what merely is, has been, or will be."[30]

The term "voluntary" implies that ethics' purpose is not only understanding what is right and what is wrong, but must also provide motivations to do it. In other words, it needs not only to state what is right, it needs not only to state why something is right, but needs also to state why someone should *do* what she sees to be right. The motivation given by ethics must be rational. In Sidgwick's words, "I think that when a man seriously asks 'why he should do' anything, he commonly assumes in himself a determination to pursue whatever conduct may be shown by

[26] For example, the Concluding Chapter is here partly summarized when discussing the third book and partly when discussing the fourth book, but never as itself.
[27] (Sidgwick, 1874, p. 1)
[28] (Sidgwick, 1874, p. 4)
[29] (Sidgwick, 1874, p. 1)
[30] Ibid.

argument to be reasonable, even though it be very different from that to which his non-rational inclinations may prompt."[31]

The terms "ought" and "right" are clearly the core of the definition of Ethics. In fact, the differences between the three Methods of Ethics depend, mainly, on what is intended for "ought" and "right" in them. The differences between these ways for intending these terms come from two distinctions.

The first is consequent to the question: are terms such as "ought" and "right" always implicitly relative to an optional end? Or are they not? For example, "A physician assumes that his patient wants health: he tells him that he ought to rise early, to live plainly, to take hard exercise. If the patient deliberately prefers ease and good living to health, the physician's precepts fall to the ground: they are no longer addressed to him."[32] The term "ought" might not have been explicitly stated by the physician as relative to the goal of good health, but is tacitly presupposed as such. The tacit presupposition of an end is easily recognized in cases as the one just mentioned, but it is not so easily recognizable in sentences such as "it is not right to hurt another human being for no reason". In fact, according to one of the Methods, in sentences like the one just stated the implicit presupposition of an end does not exist. As Sidgwick himself recognized, the matter is if one admits the existence of what Kant calls Categorical Imperatives or if admits only the existence of Hypothetical Imperatives.

A Method that does not allow that "ought" and "right" are always relative to optional ends, which means that the Categorical Imperative is hold to exist, is an Intuitional method. This Method is called Intuitional, since, according to it, "conduct is held to be right when conformed to certain precepts or principles of Duty, *intuitively known* to be unconditionally binding".[33]

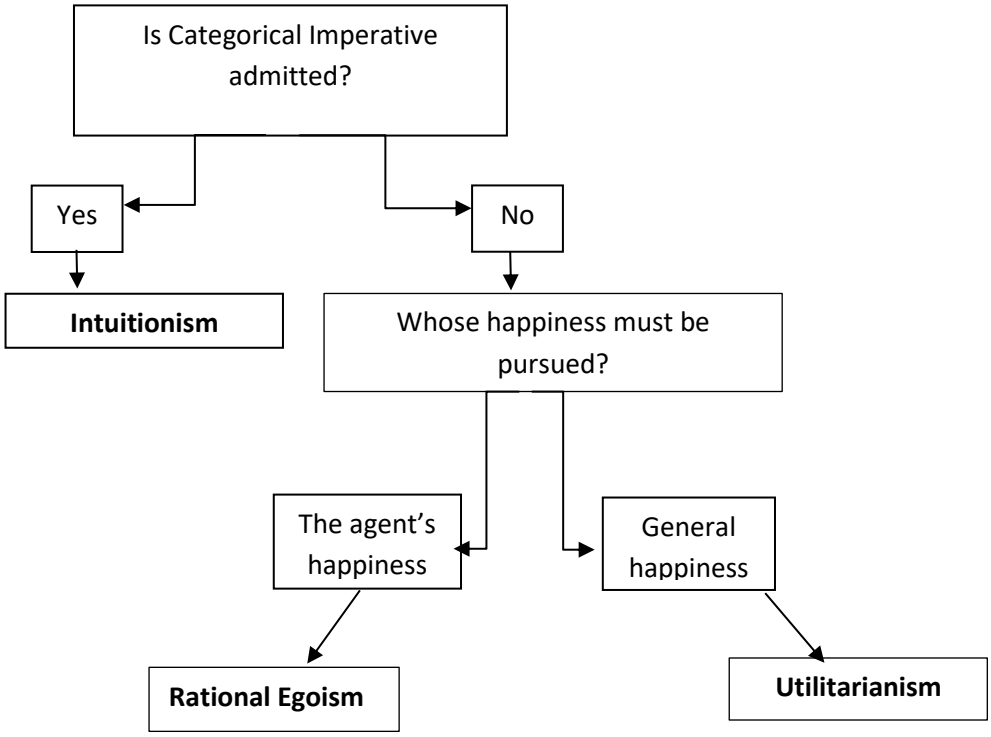
A Method that presumes that "ought" and "right" have no meaning if not referred to an end, which means that only the Hypothetical Imperative is believed to exists, is Utilitarian or Egoist. I shall explain the difference between them soon. What are terms such as "ought" and "right" always implicitly referred to, according

[31] (Sidgwick, 1874, p. 5) The relationship between moral reasons and non-rational desires will be further developed by Nagel, as shown in chapter 5.
[32] (Sidgwick, 1874, p. 7)
[33] (Sidgwick, 1874, p. 3)

to the Utilitarian and Egoist methods? These methods suppose that an agent "ought" do something, or her action is "right", when it leads to happiness,[34] which is identified as the greatest surplus of pleasure over pain.

The second distinction needed in order to comprehend Sidgwick's identification of the three Methods is between those who hold that the happiness to be pursued must be the one of the agent himself (this position is called Rational Egoism, or simply Egoism), and those who hold that the ultimate aim for action must be the general happiness, not happiness for an individual (Utilitarianism). Sidgwick notices that "whereas the philosopher seeks unity of principle, and consistency of method at the risk of paradox, the unphilosophic man is apt to hold different principles at once, and to apply different methods in more or less confused combination."[35]

We can schematize what just said as follows:



[34] What I write is a simplification, but complicate this matter is not necessary here. Actually Sidgwick distinguishes between those who hold the supreme aim for moral acting Happiness and those who hold as supreme aim for moral acting the perfection, or excellence of human nature. In his treatise Sidgwick shows how this latter view actually admits the Categorical Imperative and must consequently be considered Intuitional. I will not examine Sidgwick's reasoning on this matter, because that would lead us astray. I only inform that Sidgwick's argument can be found in (Sidgwick, 1874, p. 391-407).

[35] (Sidgwick, 1874, p. 6)

1.3. Intuitionism

Sidgwick defines Intuitionism as follows:

> *Intuitionism:* "the view of ethics which regards as the practically ultimate end of moral actions their conformity to certain rules or dictates of Duty unconditionally prescribed."[36]

According the intuitionist view how does a moral agent manage in accordance with the Categorical Imperatives, which are "rules or dictates of Duty unconditionally prescribed"? How does she acknowledge them? According to Sidgwick, there are three views on this matter. One may acknowledge the Categorical Imperatives:

1) by appealing to his conscience, which is "the accepted popular term for the faculty of moral judgment, as applied to the acts and motives of the person judging".[37] This faculty needs not to know the general rules: conscience needs only to be appealed and will suggest the right conduct. This view may be seen as too extreme, since it might recognize as decisive "simple immediate intuitions alone and discards as superfluous all modes of reasoning to moral conclusions".[38] Sidgwick calls this view Perceptional Intuitionism.

2) by appealing to an authority. What conscience suggests is not always indubitable and clear. Nevertheless, some general rules exist and are discernable; only, understanding them requires an exceptional sensibility or practice. Thus, when in doubt, we must rely on a moralist, whose function "is to perform this process of abstract contemplation, to arrange the results as systematically as possible, and by proper definitions and explanations to remove vagueness and prevent conflict."[39] Sidgwick calls this view Dogmatic Intuitionism.

---

[36] (Sidgwick, 1874, p. 96)
[37] (Sidgwick, 1874, p. 99)
[38] (Sidgwick, 1874, p. 100)
[39] (Sidgwick, 1874, p. 101). I notice that Sidgwick's analysis is here not precise. In fact, he notices (Sidgwick, 1874, p. 100) how Christians (and thus, I think, believers of some other religion) have to be classified as Dogmatic Intuitionists, which I find correct, but conflicting with his description of the role of the moralist. I think that Christians (and believers of some other religion) appeal to the authority of a sacred Book. If they rely on an exegete in order to understand its precepts, Sidgwick's definition of the moralist may work, even if it must be noticed that the "abstract contemplation" is based on the sacred Book rather than being completely abstract. If otherwise they interpret the sacred Book by themselves, the moralist is not a person but, it might be stated, the sacred Book itself, from

3) by reasoning. The morality of common-sense, which is what conscience suggest us, "even when made as precise and orderly as possible, is often found unsatisfactory as a system"[40] even if its general authority is not in doubt. "Even granting that these rules can be so defined as perfectly to fit together and cover the whole field of human conduct, without coming into conflict and without leaving any practical questions unanswered, still the resulting code seems an accidental aggregate of precepts, which stands in need of some rational synthesis. [. . .] we may yet require some deeper explanation *why* it is so"[41] by attempting to find its philosophical basis. This view is called Philosophical Intuitionism.

Perceptional Intuitionism seems philosophically unreliable, as explained when defining the other two, and some forms of Dogmatic Intuitionism seems to require Philosophical Intuitionism. Sidgwick dedicates the third book of his Methods mainly to Philosophical Intuitionism. Further details on Intuitionism will be given when analyzing this book.

1.4. Pleasure and Pain

As we have seen, according to the Methods of Rational Egoism and Utilitarianism, terms such as "ought" and "right" are implicitly referred to the obtainment of happiness. I also stated that happiness is intended as the greatest surplus of pleasure over pain. This latter statement, which can be found unsatisfying by someone, needs clarification. Then, in order to understand Rational Egoism and Utilitarianism, *Pleasure* and *Pain* have to be defined.

According to Sidgwick happiness is defined "as convertible with Pleasure, or rather as denoting that of which the constituents are pleasures", considering the term Pleasure not in the sense of pure exaltation and enjoyment, but "in its widest sense, as including every species of 'delight,' 'enjoyment,' or 'satisfaction'; except so far as any particular species may be excluded by its incompatibility with some greater pleasures, or as necessarily involving concomitant or subsequent pains."[42]

---

which the believers try to find clearness on what conscience suggests. In this case, there is no systematic arrangement of any kind, and once again the "abstract contemplation" is based on the sacred Book.

[40] (Sidgwick, 1874, p. 102)

[41] Ibid.

[42] (Sidgwick, 1874, p. 93)

How is *Pleasure* to be defined, then? Several thinkers tried to answer this question, and Sidgwick analyzes a great number of those answers.[43] He seeks a good definition for measurement purpose, because, as we will see soon, pleasure and pain must be measured in order to decide what the best action to perform is. Sidgwick concludes that the most complete definition, and most useful for measurement purpose, is the following:

> *Pleasure*: "feeling apprehended as desirable[44] by the sentient individual at the time of feeling it". [45]

Consequently, we may define *Pain* as its negative, or:

> *Pain*: feeling apprehended as not desirable by the sentient individual at the time of feeling it.

I said that, according to a method of Ethics that does not allow Categorical Imperatives, we ought to act in order to obtain the greater amount of pleasure. But is this a reliable criterion? J.S. Mill wrote[46] what would have been an objection to Sidgwick's view. He wrote that considering only the *quantity* of pleasure is a too approximate view, and that also *quality* has to be taken into account: there are pleasures whose quality is so higher than others' that no amount of the latter will be preferable to the former. According to Sidgwick, this is not decisively true; what Mill notices is not enough for stating that we should consider anything else than the *quantity*. In fact, if something gives a better quality of pleasure, it only means that it

[43] (Sidgwick, 1874, p. 125-130)
[44] One might ask the definition of Desire. Sidgwick defines it as "a felt impulse or stimulus to actions tending to the realization of what is desired." (Sidgwick, 1874, p. 43). So the term "desirable" might be defined as something that causes a felt impulse or stimulus to actions tending to the realization of that 'something'. Pleasure is desirable "when considered merely as a feeling and not in respect of its objective conditions or consequences, or of any facts that come directly within the cognizance and judgement of others besides the sentient individual" (Sidgwick, 1874, p. 131). It must be noticed that Sidgwick's definition of Pleasure does not imply that the stronger the desire, the higher the pleasure. Neither is implied that the object of our desire is always pleasure. About the relationship between Pleasure and Desire Sidgwick wrote a chapter, (Sidgwick, 1874, p. 39-56), that is relevant for Utilitarian history but for the greatest part is not relevant for the present work. The only relevant exception is the Fundamental Paradox of Hedonism, that is summarized at the end of this examination of Sidgwick's concepts of *Pain* and *Pleasure.*
[45] (Sidgwick, 1874, p. xxviii) and (Sidgwick, 1874, p. 131), and better explained in (Sidgwick, 1874, p. 125-130)
[46] (Mill, 1861, p. 183-202)

causes a higher degree of pleasure. But the degree of pleasure caused by something is, in fact, a quantity. In other words, Sidgwick admits, as Bentham, that some pleasures are "impure", which means that obtaining them cannot happen without obtaining also some pain, but this does not mean that we must consider other than the total quantity of pleasure.

It is also important, at least for measurement purpose, to discern between intensity of *pleasure* and intensity of *sensation*, "as a pleasant feeling may be strong and absorbing, and yet not so pleasant as another that is more subtle and delicate."[47]

How does someone understand which one is the act that leads to the greatest amount of pleasure? She does it by comparison of the consequences of possible acts. Egoism and Utilitarianism assumes that, for any list of possible actions, those actions can be arranged in a scale according to their total of generated pleasure or pain. Generated pleasure would be a positive value, which makes the choice more appealing, and generated pain a negative value, that makes the choice less appealing.

Sidgwick notices that this leads us "to the assumption of a hedonistic zero, or perfectly neutral feeling, as a point from which the positive quantity of pleasures may be measured." The idea of the "hedonistic zero", as some other concepts from *The Methods of Ethics*, is an idea whose importance was ignored by the majority of thinkers until Parfit's *Reasons and Persons*, in the fourth and last part of which the problem of establishing what is the "hedonistic zero", and if it is desirable, becomes extremely relevant. More particularly, the concept of Hedonistic Zero raises puzzling questions in the field that the fourth part of *Reasons and Persons* initiates, which is called Population Ethics. Population Ethics will be the focus of the third part of this work, and will be better defined there; for the moment I only state that it focuses, roughly, on establishing how many people should exist. Sidgwick describes the hedonistic zero as a

> perfectly neutral feeling, [. . .] a point from which the positive quantity of pleasures may be measured. And this latter assumption [according to which a "hedonistic zero" has to exist] emerges still more clearly when we consider the comparison and balancing of pleasures with pains [. . .]. For pain must be reckoned as the negative quantity of pleasure, to be balanced against and subtracted from the positive in estimating happiness on the whole; we must

[47] (Sidgwick, 1874, p. 94)

> therefore conceive, as at least ideally possible, a point of transition in consciousness at which we pass from the positive to the negative.[48]

Must we think that the "hedonistic zero" is purely ideal, an abstract and extremely theoretical concept, and is not a feeling ever experienced by anyone? Sidgwick does not think so; he writes "experience seems to show that a state at any rate very nearly approximating to it is even common: and we certainly experience continual transition from pleasure to pain and vice versa, and thus (unless we conceive all such transitions to be abrupt) we must exist at least momentarily in this neutral state."[49] The fact that Sidgwick admits a "hedonistic zero" in our experience does not imply that he holds that the "hedonistic zero" is the normal condition of our consciousness,

> out of which we occasionally sink into pain, and occasionally rise into pleasure. Nature has not been so niggardly to man as this: so long as health is retained, and pain and irksome toil banished, the mere performance of the ordinary habitual functions of life is, according to my experience, a frequent source of moderate pleasures, alternating rapidly with states nearly or quite indifferent. Thus we may venture to say that the 'apathy' which so large a proportion of Greek moralists in the post-Aristotelian period regarded as the ideal state of existence, was not really conceived by them as 'without one pleasure and without one pain'; but rather as a state of placid intellectual contemplation, which in philosophic minds might easily reach a high degree of pleasure.[50]

This means that, according to Sidgwick, life is, generally, worth living. As we will see in 4.1, this has some implications in some of his beliefs about what can be described as the first recorded reasoning about Population Ethics.

Sidgwick writes that a method that does not allow Categorical Imperatives must decide which one is the best choice by comparing different acts basing on their pleasantness. There are many arguments against the possibility of such a comparison.[51] For example, it may be argued (1) that the act of observing and examining pleasure limits the full and intense experience of pleasure itself. Since it is clear that we cannot compare anything if we do not observe it, this objection may

[48] (Sidgwick, 1874, p. 124)
[49] (Sidgwick, 1874, p. 124-125)
[50] (Sidgwick, 1874, p. 125)
[51] This theme, as some other I mentioned in this section, is analyzed in the second book of *The Methods of Ethics*, but I think it is the case to anticipate its analysis for clearness' sake.

have some force.[52] In addition it might be claimed (2) that this actual comparison is, in an ordinary life, both occasional and very rough, and in any case it is "liable to illusion, of which we can never measure the precise amount, while we are continually forced to recognise its existence."[53] Or again (3), there is no way to tell that a past pleasure will satisfy us in the future as it did previously, or will be enjoyed by others[54] as it is enjoyed by us.[55]

Sidgwick rejects the first objection and accepts partly the other two, even if he thinks that they are not decisive. He answers to (1) in this way: even if it might be true that observation of pleasure and its enjoyment are incompatible (which is sometimes false), the mere consciousness of a present feeling suffices for comparison. This is so because we can enjoy pleasure without examining it, but examine it later when we have to compare it with other pleasure. Thus, examination and enjoyment are compatible if the examination does not occur at the time of enjoyment, which is usually the case.

In order to examine objection (2), on the roughness of comparison of our pleasures, Sidgwick tries to explain how we estimate the value of different pleasures for practical purposes. It is worth reporting entirely his words. According to Sidgwick, the process is the following:

> We project ourselves into the future, and imagine what such and such a pleasure will amount to under hypothetical conditions. This imagination, so far as it involves conscious inference, seems to be chiefly determined by our own experience of past pleasures, which are usually recalled generically, or in large aggregates, though sometimes particular instances of important single pleasures occur to us as definitely remembered: but partly, too, we are influenced by the experience of others sympathetically appropriated: and here again we

[52] This objection and its answer, which I will summarize later, can be found in (Sidgwick, 1874, p. 138-40).

[53] (Sidgwick, 1874, p. 140) . This objection and its examination, which I will summarize later, can be found in (Sidgwick, 1874, p. 140-144).

[54] As we will see in chapter 7, according to Parfit's theory of personal identity, the sentence "there is no way to tell that a past pleasure will satisfy us in the future as it did previously" and "there is no way to tell that a past pleasure will be enjoyed by others as by us at the present time" are equivalent, at least at some extent.

[55] This objection and its examination, which I will summarize later, can be found in (Sidgwick, 1874, p. 147-150). In (Sidgwick, 1874, p. 131-150) Sidgwick comments other objections in addition to the three I have reported here. I chose to analyze only these three because the first seems a philosophically interesting question whose answer may have some relevance, and the other two because they helps us understand a limit of Egoism and Utilitarianism. For other objections, see (Sidgwick, 1874, p. 131-150).

> sometimes definitely refer to particular experiences which have been communicated to us by individuals, and sometimes to the traditional generalisations which are thought to represent the common experience of mankind.[56]

Clearly, this system of confrontation is not perfect: mistakes can occur. Also, it seems really hard to compare really different pleasures, such as an intense intellectual pleasure and an intense physical pleasure. Furthermore, when someone compares the same pleasure in different situations she judges it in different ways: for example the pleasure of a physical activity may appear more agreeable when we are well-rested than when we are very tired. Thus, measurement cannot be precise if there is no other way for comparing them than basing on empirical knowledge.

For similar reasons, the objection (3) is effective to some extent. Past pleasure may not satisfy in future,

> For our capacity for particular pleasures may be about to change, or may have actually changed since the experiences that form the data of our calculation. We may have reached the point of satiety in respect of some of our past pleasures, or otherwise lost our susceptibility to them, owing to latent changes in our constitution: or we may have increased our susceptibility to pains inevitably connected with them: or altered conditions of life may have generated in us new desires and aversions, and given relative importance to new sources of happiness.[57]

Neither we can be sure that what we enjoy will be pleasurable for others,

> For such inference proceeds on the assumption of a similarity of nature among human beings, which is never exactly true, while we can never exactly know how much it falls short of the truth; though we have sufficient evidence of the striking differences between the feelings produced in different men by similar causes, to convince us that the assumption would in many cases be wholly misleading.[58]

We can never know exactly someone else's susceptibility to different kinds of pleasure and pain, and how much their susceptibility is similar to or different from

[56] (Sidgwick, 1874, p. 141-142)
[57] (Sidgwick, 1874, p. 147)
[58] (Sidgwick, 1874, p. 148)

ours. Furthermore, there are some pleasures whose enjoyment requires a feature that is not universal between individuals, neither is present in someone's life from the beginning; for example "the sacrifice of sensual inclination to duty is disagreeable to the non-moral man when he at first attempts it, but affords to the truly virtuous man a deep and strong delight. And similarly almost all the more refined intellectual and emotional pleasures require training and culture in order to be enjoyed".[59] Since these enjoyments exist, we cannot be sure that others, or ourselves in the future, will have those features in the same degree that let us enjoy a particular pleasure at present time.

If Sidgwick admits that those two objections reduce our confidence on the comparison between pleasures, at least if based on reflection of empirical experiences, why does he not think that they are decisive? He writes:

> I am conscious that, in spite of all the difficulties that I have urged, I continue to make comparisons between pleasures and pains with practical reliance on their results. But I conclude that it would be at least highly desirable, with a view to the systematic direction of conduct, to control and supplement the results of such comparisons by the assistance of some other method: if we can find any on which we see reason to rely.[60]

In other words, we cannot help but compare them, even if this comparison is not precise; and, since we do not currently have another method for comparison,[61] for now we must base the comparison on empirical experiences.[62]

---

[59] (Sidgwick, 1874, p. 150)
[60] Ibid.
[61] According to Sidgwick, human beings have little hope to obtain useful methods for comparison, at least from science. Sidgwick does not think likeable, for example, that science will ever be able to compare physical pain and pleasure with psychical pain and pleasure. See (Sidgwick, 1874, p. 176-195) for his full argument on this matter. To be fair, in the over 130 years that split us from Sidgwick, researches about this matter seem not to have done any decisive step forward yet: see for example (Kwon, 2016).
[62] Sidgwick says in (Sidgwick, 1874, p. vii) that he studied "the Theory of Evolution in its application to practice". He certainly studied the Theory of Evolution of his age, and did not link pleasure to evolution that much. Today I believe that a further objection to empirical estimation of pleasures and pains can be done basing on the Theory of Evolution. Some evolutionists, such as Richard Dawkins (see for example (Dawkins, 2006)), or behaviorist psychologists such as Howard Rachlin (see for example (Rachlin, 2010)), may suggest that pleasure is a reward that an organism gives himself in order to be encouraged to repeat a certain action. For example, eating is a pleasure for our organism only because organisms with stronger motivations for eating eat more frequently; since eating is helpful for survival, an organism that receives pleasure for eating (and thus eats frequently) has more probability of survival than an organism that does not. I think that this does not mean that we have pleasure or pain only in order to survive, since we experience pleasure also when our survival is not

But even admitting that the comparison between pleasures is possible, Sidgwick needs to address another, more challenging problem: seeking pleasure in itself is *not* an effective way to obtain pleasure. This sentence, although paradoxical at first sight, is actually common knowledge: in many activities it is more rewarding the pursue of an objective than its attainment.

> Take, for example, the case of any game which involves as most games do a contest for victory. No ordinary player before entering on such a contest, has any desire for victory in it: indeed he often finds it difficult to imagine himself deriving gratification from such victory, before he has actually engaged in the competition. What he deliberately, before the game begins, desires is not victory, but the pleasant excitement of the struggle for it; only for the full development of this pleasure a transient desire to win the game is generally indispensable.[63]

Another example is the pleasure given by scientific research: the researcher clearly is not after pleasure, but after some kind of knowledge. Countless other examples may be added to support the objection to non-Intuitionist methods of ethics according to which "it would be a palpable mistake to say that this prospective pleasure is the object of the desire that causes it".[64] It is extremely hard to deny that, if someone merely focuses on obtaining pleasure, she will not obtain it, at least not at its fullness. The objection is quite powerful: if aiming at pleasure does not give pleasure, theories

---

directly in danger, and sometimes even when pleasure is caused by actions that damage our organism rather than conserving it, for example eating certain unhealthy food. This might mean that natural selection leads us not only to survive, but also to survive at certain conditions of "well-being", that might be connected with life-preservation but do not aim precisely at it. For example, friend's company is pleasurable, might be linked somehow with self-survival but cannot be satisfactorily described as aiming only to survival. It seems that our organism tends not only to live, but also to live "well", or at least tends to live in certain conditions generally regarded as agreeable. If pleasure aims at the obtainment of self-preservation or some satisfactory conditions of self-preservation, we might need to compare not pain and pleasure, but to compare how successfully we will achieve self-preservation, its satisfactory conditions or whatever pain and pleasure should push us to search due to the evolutionistic selection. Certainly, pain and pleasure are a good clue to understand which choice will lead to the most successful obtainment of self-preservation and its satisfactory conditions, but they are not the only means for understanding the best choice, since reasoning might help (which no Egoist and no Utilitarian thinker ever denied), and most importantly they are not the final aim. Thus, as it is for objections (2) and (3) of this part, I do not think that this objection can be ignored or denied, if the rightness of evolutionists like Dawkins or psychologists like Rachlin have to be admitted. However I do not think it can be regarded as decisive for undermining Egoism and Utilitarian thinking, since the method of comparison can certainly be corrected by reason, but cannot disregard empirical experience, whose importance remains paramount.

[63] (Sidgwick, 1874, p.47)

[64] Ibid.

as utilitarianism and Rational Egoism, that prescribes the obtainment of pleasure, are inapplicable. This objection is so strong that Sidgwick decides to call it

> *The Fundamental Paradox of Hedonism:* "that the impulse towards pleasure, if too predominant, defeats its own aim".

The paradox is a stronger objection to Egoism than to Utilitarianism: actions that bring pleasure to others seems less affected by this paradox. Egoism does not consider altruist actions as rational, therefore is more affected by the paradox than utilitarianism, that allows altruist actions.

In order to solve the paradox, Sidgwick has to do something that seems conceptually impossible: he needs to show that a theory that defeats itself may nevertheless be a reliable theory. The astonishingly brilliant solution is that a self-defeating theory may work if it is *self-effacing.*[65] In Sidgwick's words:

> The principle of Egoistic Hedonism, when applied with a due knowledge of the laws of human nature, is practically self-limiting; i.e. that a rational method of attaining the end at which it aims requires that we should to some extent put it out of sight and not directly aim at it. I have before spoken of this conclusion as the 'Fundamental Paradox of Egoistic Hedonism'; but though it presents itself as a paradox, there does not seem to be any difficulty in its practical realisation [. . .]. For it is an experience only too common among men, in whatever pursuit they may be engaged, that they let the original object and goal of their efforts pass out of view, and come to regard the means to this end as ends in themselves: so that they at last even sacrifice the original end to the attainment of what is only secondarily and derivatively desirable. And if it be thus easy and common to forget the end in the means overmuch, there seems no reason why it should be difficult to do it to the extent that Rational Egoism prescribes: and, in fact, it seems to be continually done by ordinary persons in the case of amusements and pastimes of all kinds.[66]

[65] The term is Parfit's. The importance of Sidgwick's solution of this Paradox, as the concept of hedonistic zero, has not been comprehended in its fullness until *Reasons and Persons*. Part one of Reasons and Persons is focused on an analysis of self-defeating and self-effacing theories. An interesting result of this analysis is that the Morality of Common Sense, if not revised, is sometimes self-defeating too (See (Parfit, 1984, p. 95-114)). Part one of *Reasons and Persons* will be commented on chapter 6. Why Morality of Common Sense is self-defeating will not be reported, since it is not central for our work.

[66] (Sidgwick, 1875, p. 137)

1.5. Reason

Sidgwick cannot avoid examining the relationship between reason and morality: if there is no such relationship, then his work would be meaningless.

I state more clearly the matter: Hume, and many who agree with him, thinks that "Reason, meaning the judgement of truth and falsehood, can never of itself be any motive to the Will"[67]. According to Hume, only desires, that are irrational, can provide motivation: according to him, people would be motivated only something irrational, and not by Reason. If this is true, what would the relevance of examining Methods of Ethics be, given that a Method of Ethics is, as I said before, "any *rational* procedure by which we determine what individual human beings 'ought' or what it is 'right' for them to do, or to seek to realise by *voluntary* action"?[68] If Reason as itself cannot motivate will, no rational procedure can determine a voluntary action, thus any examination on the Methods of Ethics, at least as Sidgwick intends it, is meaningless. In order to prove that his work is justified, Sidgwick must demonstrate that reason as itself *can* be motivational for Will.

It is hard to deny that sometimes there is a conflict between reason and irrational or non-rational desires:[69] it is common experience and someone might think that this experience alone is enough to refuse Hume's statement. In fact, this experience seems to testify that our reason has some force in our decision, thus it may seem that Hume's view according to which we are merely governed by irrational desires is false. But Hume believes that this conflict can better be described as "merely a conflict among desires and aversions; the sole function of reason being to bring before the mind ideas of actual or possible facts, which modify [. . .] the resultant force of our various impulses".[70] Therefore experience is not enough to reject Hume's idea of Reason as non-motivating in itself.

---

[67] (Sidgwick, 1874, p. 23); this concept by Hume can be found in *A treatise on Human Nature,* Volume II, part III, Section III.

[68] (Sidgwick, 1874, p. 1)

[69] Sidgwick distinguishes "irrational" and "non rational". Irrational means "not examined in our consciousness and against what would have been our deliberate judgement", for example phobias. "Non rational" means "not examined in our consciousness but not in contrast with any deliberate judgement". For example, when a person eats her dinner she does not examine carefully in her conscience if dining is, in that moment, right, wrong, prudent or imprudent; neither dining is usually judged an act against reason. Further about this distinction can be read in (Sidgwick, 1874, p. 23-25).

[70] (Sidgwick, 1874, p. 25)

Sidgwick thinks that, in order to reject this theory by Hume, it is convenient to show the inadequacy of all attempts to explain the practical judgments in which notions such as “ought” or “right” are conceived as lacking of motivational power. If no explanation for ethical judgements will appear satisfactory unless “ought” and “right” are motivational, then the explanation according to which “ought” and “right” are motivational will result better and therefore Sidgwick’s examination will result meaningful.

The most common explanations for ethical judgement according to which “ought” or “right” are *not* motivational are (1) that a moral judgement means only that the person who judges feels aversion or attraction for an act (2) that a moral judgement means that the conduct in question is prescribed under penalties.

Explanation (1) implies, for example, that a statement like “Truth ought to be spoken” means that who made the statement felt an approbation or satisfaction excited by the idea of truthspeaking. Sidgwick admits that “probably some degree of such emotion, commonly distinguished as 'moral sentiment', ordinarily accompanies moral judgments on real cases”,[71] but describing a moral judgement as a mere subjective attraction is incorrect, “otherwise the fact of another man's disapprobation might equally be expressed by saying 'Truth ought not to be spoken'; and thus we should have two coexistent facts stated in two mutually contradictory propositions”,[72] and a man who thinks that truth ought to be spoken would not find nothing surprising in this contradiction, since in all judgements regarding taste it is admitted that what someone likes can be disliked by someone else. But when someone says that “truth ought to be spoken” he admits no opposite statement to be believed true: she does not think their judgement as subjective, and thinks that no one would disagree with his judgement. Thus explanation (1) is not satisfactory.

Explanation (2), according to which a moral judgement means that the conduct in question is prescribed under penalties, leads to ask who prescribed that conduct, and thus who will inflict the penalty. There are three main answers to this question.

According to one answer prescriptions and punishments are done by human Law; this answer is not satisfying, because otherwise we could not conceive sentences like

[71] (Sidgwick, 1874, p. 27)
[72] Ibid.

"Law must be Equal", since it is not commonly intended with this sentence that Law has prescribed to itself to be equal, and will punish itself if it will not be. According to another answer, public opinion is the author of prescriptions, and punishment consists in social disapprobation. We surely fear the judgement of our surroundings, but certainly Sidgwick is right when he says that "there are many things which we judge men 'ought' to do, while perfectly aware that they will incur no serious social penalties for omitting them".[73] For example, is commonly accepted that everyone ought to discuss every argument calmly and trying to comprehend who disagrees with her, but if someone gets excited and speaks louder than usual against someone that has different ideas, refusing to listen to other opinions rather than her own, this speaker is generally not significantly disapproved by public opinion. Therefore, we cannot be satisfied stating that public opinion prescribes conduct and penalties.

The third hypothesized author of conduct and punishment, being the first one human law and the second one public opinion, is Divine Law. Someone ought to do something because God prescribed so. If she does not do what they ought to do, they will be punished in the afterlife. According to Sidgwick, this view is not shared much, not even between religious people.[74] If someone holds this view, she is mistaken, because religious people believe that God will punish wrong actions because He is *right* to punish them; in other words, wrong action *ought* to be punished, and He is the one who has to; this cannot mean that God bounds himself to his own punishment, and that he will punish himself if he does not punish others. So this explanation fails.

The meaning of moral obligations seems therefore to be something more than a subjective approbation in the judger's mind and more than the fact that certain rules of conduct are supported by penalties which will follow on their violation. Sidgwick writes:

> What then, it may be asked, does it import? What definition can we give of 'ought,' 'right,' and other terms expressing the same fundamental notion? To this I should answer that the notion which these terms have in common is too elementary to admit of any formal

[73] (Sidgwick, 1874, p. 29)
[74] (Sidgwick, 1874, p. 31)

> definition. [. . .] A psychologist must accept as elementary what introspection carefully performed declares to be so; and, using this criterion, I find that the notion we have been examining, as it now exists in our thought, cannot be resolved into any more simple notions.[75]

Sidgwick's analysis proceeds noticing that the word "ought" can have two meanings. According to the narrower, "what we judge 'ought to be' done, is always thought capable of being brought about by the volition of any individual to whom the judgment applies. I cannot conceive that I 'ought' to do anything which at the same time I judge that I cannot do".[76] According to the wider meaning "I 'ought' to know what a wiser man would know, or feel as a better man would feel, in my place, though I may know that I could not directly produce in myself such knowledge or feeling by any effort of will. In this case the word merely implies an ideal or pattern which I 'ought' in the stricter sense to seek to imitate as far as possible."[77]

A good way to describe a term that is not composed by more essential concept, such as it is "ought", might be examining its relationship with other ideas. Sidgwick observes that both in the narrow and in the wide meaning is implied "that what ought to be is a possible object of knowledge: *i.e*. that what I judge ought to be must, unless I am in error, be similarly judged by all rational beings who judge truly of the matter."[78] Rational beings are able to make such judgements due to a faculty that Sidgwick calls Reason. It is defined as such:

> *Reason*: "faculty of moral cognition"[79]

It must be noticed that Sidgwick does not clarify what are the relationship between the concept of Reason just stated and Hume's concept of Reason, according to which Reason can only distinguish true and false. Sidgwick never says[80] that the faculty of moral cognition *also* distinguishes right and wrong, but a logical reasoning seems implied in all Methods of ethics but Perceptional Intuitionism. Thus it seems unlikely

[75] (Sidgwick, 1874, p. 32-33)
[76] (Sidgwick, 1874, p. 33)
[77] Ibid.
[78] Ibid.
[79] (Sidgwick, 1874, p. 34)
[80] He neither denies it, to be fair.

that Sidgwick rejects the idea that a faculty for distinguishing true and false is required in morality, since moral reasoning is frequently intended as supported by the faculty of distinguishing true and false. We cannot help but wonder if this epistemic faculty for distinguishing true and false is different from a moral faculty that lets us discern logically right and wrong, or both of them are a single faculty, whose only task is to logically examine possible actions, information or else. If the two faculties are to be distinguished, the epistemic faculty would be an important instrument for the correct functioning of the Moral faculty, whereas the moral faculty seems to be of no use in the epistemic faculty.[81]

Sidgwick may be unclear about the relationship between his concept of Reason and Hume's, but states clearly that when any rational being recognizes something that "ought to be done" it is implied that this recognition gives to anyone a motive or an impulse to do it. Sidgwick cannot deny that this motive will be one among many others, and that there is no way to predict if those other motives will be stronger than the moral motive and therefore the moral action will not be done. Nevertheless, Reason faculty, whatever the relationship between Epistemic and Moral reason are, judges right and wrong, and this judgement gives reasons to act. [82]

The term Reason has another meaning, that is, roughly, "rational motive", as in the sentence "there is no reason to perform this act". The concept of Reason as such will be the core of our later analysis of Nagel's *The Possibility of Altruism* and will have a paramount role in Nagel's attempt to resolve the Dualism of Practical Reason. But does this Dualism really exist?

---

[81] I think that the two faculties are probably to be distinguished. Sidgwick's definition of Reason implies that reason gives some motives. An Epistemic faculty of Reason, like the one described by Hume, does not gives motives: no motive for action comes from the correctness of 2+2=4 in itself, or from the wrongness of 15+18=36 in itself. If Sidgwick intends the two faculties as separated, in my opinion, he would not be in contraposition with Hume about the concept of reason: simply, Hume calls Benevolence the faculty of moral cognition (about Benevolence see (Hume, 1751)), and calls Reason something else, that Sidgwick never examines in his book. Obviously I am not stating that Hume's theory of Benevolence is perfectly equivalent to Sidgwick's concept of Reason. Sidgwick's Reason is a *formal* faculty: it might be a feeling corrected by reason in its effects, like Hume's Benevolence is, it might be something else. It is not Sidgwick's concern describing it: he ascertains its existence, its relation with moral statements, and its implication of being motivational. Therefore, if Sidgwick distinguishes between epistemic and moral reason, neither Hume's and Sidgwick's concept of Reason are in contraposition, since they focus on different matters, nor Hume's benevolence and Sidgwick's reason can be described as in contrast.

[82] Clearly Sidgwick admits that this faculty can fail. If I have reasons to do something, it means that my moral faculty judges that this act is morally right, but its judgement can be wrong, if, for example, my moral faculty did not consider a data of the situation in which I am acting, or has done some logical error.

Chapter 2: Egoism

Egoistic Hedonism is described by Sidgwick as the Method of Ethics that prescribes the "adoption of his own greatest happiness as the ultimate end of each individual's actions."[83] In the previous chapter several features of this method has been analyzed, and several objections, such as the Fundamental Problem of Hedonism, have been addressed. We need now to focus on the following questions: how solid are the basis of Sidgwick's Profoundest Problem? Is it true that there is no link between self-interest and duty? What are the relationships between happiness and altruism, if there is any?

As a matter of fact, according to many thinkers, being a virtuous man is highly desirable because it is advantageous. This is not an old dead opinion that can be found only in some ancient Greek philosopher, for example Plato,[84] only to disappear after the middle-age. There are some utilitarian philosophers[85] who would agree with Plato: someone that is a good person receives benefits from it. Sidgwick summarizes this position with those words:

> [According to some utilitarian authors] Virtues [. . .] are qualities either useful or directly agreeable to others: thus they either increase the market value of the virtuous man's services, and cause others to purchase them at a higher rate and to allot to him more dignified and interesting functions; or they dispose men to please him, both out of gratitude and in order to enjoy the pleasures of his society in return: and again since man is an imitative animal the exhibition of these qualities is naturally rewarded by a reciprocal manifestation of them on the part of others, through the mere influence of example.[86]

Is this true? And is it enough for stating that the Profoundest Problem does not exist? Sidgwick addresses the problem in two ways.

One is simply observing that, in order to obtain those advantages, someone needs not to observe duty: it will suffice *appear* to be useful to others. Furthermore, a

[83] (Sidgwick, 1875, p. 119)
[84] See for example Plato's *Republic.* I clearly am not stating that Plato's opinion are usually old and dead.
[85] Sidgwick does not specify who he is referring to. In my opinion, an example of this kind of thinking can be found in (Hume, 1751), especially in section V, and in the section I of part III of the third book of Hume's *Treatise on Human nature*.
[86] (Sidgwick, 1874, p. 167)

man may need to be useful only at the eyes of those who can repay him.

The second answer is a more complete analysis of the process of approval and disapproval. His analysis begins distinguishing duties into self-regarding and public, both established by moral rules. Approval due to conformity to those rules and disapproval due to not conformity to those rules is called by Sidgwick the "sanction" of those rules. Sanction can be External, for public duties, or Internal, for self-regarding duties. The former class is further divided between legal sanctions, which are sanctions done by a recognized authority, and social sanctions, "which are either the pleasures that may be expected from the approval and goodwill of our fellow-men generally, and the services that they will be prompted to render both by this goodwill and by their appreciation of the usefulness of good conduct, or the annoyance and losses that are to be feared from their distrust and dislike."[87] Internal sanctions are the agent's emotions caused by her own virtuous behavior, such as presence or absence of remorse, or "some effect on the mental constitution of the agent produced by the maintenance of virtuous dispositions and habits."[88]

Once classified the possible sources of sanctions, Sidgwick underlines how those sanctions conflict with each other. In fact,

> The Positive Morality [which is the basis of social sanctions] of any community undergoes development, and is thus subject to changes which affect the consciences of the few before they are accepted by the many; so that the rules at any time sustained by the strongest social sanctions may not only fall short of, but even clash with, the intuitions of those members of the community who have most moral insight.[89]

This is enough for state that external sanctions conflict with internal sanction. But external sanctions conflicts also in themselves, since

> For similar reasons [similar to the reason for which external and internal sanctions conflicts with each other] Law and Positive Morality may be at variance, in details. [. . .] the more stress we lay on either the legal or the social sanctions of moral conduct, the greater difficulty

[87] (Sidgwick, 1875, p. 164)
[88] Ibid. As the reader can notice, in the internal sanctions are not included anticipations for rewards in the afterlife, that are believed by some religious people. Sidgwick here focuses explicitly only on calculations based on feelings that are not beyond the range of experience, as is stated in (Sidgwick, 1875, p 170)
[89] (Sidgwick, 1875, p. 164)

> we shall have in proving the coincidence of duty and self-interest in the exceptional cases in which we find these sanctions arrayed against what we conceive to be duty. [90]

Should a Rational Egoist, then, rely only on one between internal, legal or social sanctions?

Let's first consider if legal sanctions are enough for oblige the Egoist to act altruistically. Legal sanctions will be enough only if, by disobeying the law, the Rational Egoist goes against her own interests. If a Rational Egoist goes against the law, she will be punished by the authority, and therefore her act will be against her own interest. But not always: in fact, the subversion of law itself may give advantages to the Rational Egoist, and if the subversion is successful she will not take any damage by that. Furthermore, no legal system is perfect, consequently a Rational Egoist may calculate where to have an anti-social conduct without being punished by law. We can conclude that a Rational Egoist acting altruistically might be satisfied by legal sanctions only if this law cannot be subverted and if the legal system is perfect, but both these conditions cannot be found in existing law systems.

Let us now consider if social sanctions are enough for oblige the Egoist to act altruistically. Once again, social sanctions will be enough only if, by acting against the society's opinion, the Rational Egoist goes against her own interests. We find that considerations of fellow men is a powerful guide of conduct for an individual, and it has few flaws. Unfortunately, those flaws exist and are exactly where the defects of the legal system are. The first flaw is that the more people disagree with common opinion the weaker is social reprobation; thus a subversive behavior is not disapproved when shared by part of the community. Therefore, for example, it is not true that tyrannical usurpers or renown criminals "whose position raises them out of the reach of legal penalties"[91] suffer for the judgement of their fellow men. In addition, as law, society is not omniscient, consequently social reprobation is avoided by secret misbehaviors. As law, social sanctions are unsatisfactory to convince a Rational Egoist to act according to her duty.

The only remaining source of sanctions that may lead an Egoist to behave according to her duty are internal sanctions, which are positive emotions or effects

[90] (Sidgwick, 1875, p. 164-165)
[91] (Sidgwick, 1875, p. 166)

caused by well-doing. According to Sidgwick, it is not reasonable to presume that these positive effects on someone would reward her action so much that the reward outweighs what she denied to herself. This seems indisputable in extremes cases such as the sacrifice of a soldier, which may happen to face a "certain and painful death, under circumstances where it might be avoided with little or no loss even of reputation":[92] it is difficult to believe that emotions caused by such a conduct may outweigh the physical pain. A soldier's case might be an extreme case, but in Sidgwick's opinion a similar reasoning can be done "even in more ordinary cases, where a man is called on to give up, for virtue's sake, not life, but a considerable share of the ordinary sources of human happiness":[93] in fact, it seems difficult to argue that everyone, or that most people, is certainly repaid by their conscience for significant sacrifice done by altruist reasons. Sidgwick does not deny that there *can* be someone whose conscience is so strong that it is her interest to act altruistically, for example because the remorse for not having done a sacrifice for someone else's good would be worse than sacrifice itself; but Sidgwick believes that the majority of men is not gifted (or cursed?) with such "specially refined moral sensibilities".[94]

If Sidgwick's arguments are convincing, the discrepancy between duty and interest must be admitted and Sidgwick's Profoundest Problem is a challenge for the moral philosophers that followed him.

Most of the second book of the *Methods of Ethics* has been reported in the first chapter of this work. For our purposes we do not need to focus further on this book. After this present examination on the incompatibility between Egoism and the other two Methods, on which we will return later, we should pass on verify Sidgwick's demonstration according to which there is no contrast between Intuitionism and Utilitarianism.

Chapter 3: Intuitionism

The aim of the third book of the *Methods*, on Intuitionism, is to examine the most important rules of conduct that appear to our interior experience as a duty

[92] (Sidgwick, 1875, p. 170)
[93] (Sidgwick, 1875, p. 171)
[94] (Sidgwick, 1874, p. 175)

unconditionally binding. These rules are called moral intuitions and are the basis of any morality that might be accepted by any individual that relies mainly on what is usually called "common sense". Through examination Sidgwick tries to define and analyze moral intuitions in a scientific manner.

The book has thirteen chapters, eight of which contain a detailed analysis of the principal commonly shared principles and rules that are intuitively known as unconditionally binding and constitute the most common methods among both philosophers and common folk. The rules Sidgwick analyzes are wisdom,[95] benevolence,[96] friendship,[97] gratitude,[98] justice,[99] law observance,[100] truth observance,[101] prudence,[102] temperance,[103] purity[104] and more.[105] Sidgwick's meticulousness in this examination is remarkable, and his analysis may be said to be, if possible, sufficiently close to completeness.

Unfortunately, Sidgwick's analysis is really vast and, as the author himself admitted, to some extent "affected or pedantic".[106] In order not to unnecessarily bore

[95] (Sidgwick, 1874, p. 231-237)
[96] (Sidgwick, 1874, p. 238-263)
[97] (Sidgwick, 1874, p. 256-259)
[98] (Sidgwick, 1874, p. 259-263)
[99] (Sidgwick, 1874, p. 264-294)
[100] (Sidgwick, 1874, p. 295-311)
[101] (Sidgwick, 1874, p. 312-319)
[102] (Sidgwick, 1874, p. 327-328)
[103] (Sidgwick, 1874, p. 328-329)
[104] (Sidgwick, 1874, p. 329-331)
[105] Sidgwick examines a large number of moral principles but, curiously, he does not analyze any value such as "preservation of life" or any principle according to which it is good to preserve life. It might be because it is not universally shared: for example, the principle according to which life ought to be preserved seems to lose much of its intuitive truth on battlefields, where it might be argued that the lives of the enemy are to be destroyed. The pages in which Sidgwick stated what may be considered his opinion on the matter seems (Sidgwick, 1874, p. 414-415). From those pages, it might be concluded that he believes lives have no value in themselves, but they are still worth of the greatest consideration, since are the means to happiness. In fact, happiness is something that is experienced in life, and according to Sidgwick "generally, life on the average yields a positive balance of pleasure over pain" (Sidgwick, 1874, p. 414) and thus life ought to be preserved. This interpretation should be considered a stretch, because even if it is true that he explicitly says that a greater number of lives ought to be preferred since it implies that more people can be happy, the conclusion that according to Sidgwick lives in themselves have no value in themselves cannot be found in any part of (Sidgwick, 1874). Still, those considerations can be found in pages belong to the analysis of Utilitarianism, and not of Intuitionism. I think that an analysis of the value of life according to common sense should be made in a work like Sidgwick's one, even if only in order to state that there is not a commonly shared self-evident principle that says that life ought to be preserved. Too many people believe strongly that it is self-evident that life ought to be preserved, or at least enough people to deserve an examination in Sidgwick's treatise. Such an analysis would have been of outmost importance for Population Ethics.
[106] (Sidgwick, 1874, p. xxii) . When commenting his work, Sidgwick seems often too strict with himself: the third book may seem indeed dull to someone, for example Whitehead who, as reported in

the reader, I will summarize the analysis in the following way: I will first explore the possibility and the manners of doing a clear and precise analysis of moral intuitions; after that I will give an example of Sidgwick's analysis of a moral intuition, summarizing his examination of Justice; then I will present the difficulties shared by Philosophical Intuitionists[107] and finally I will present Sidgwick's conclusion, which is that Intuitionism is a reliable Method only if general happiness is presumed as final end of moral intuitions.

3.1. Requirements for an analysis

It is appropriate to begin an analysis on Intuitionism asking if it can be done at all. Moral intuitions are bizarre entities: according to some philosophers, they do not even exist. Do they? Furthermore, should it not be more appropriate to analyze intuitions trough an examination of intentions and motives concerning those intuitions rather than external moral standards such as prudence or humility? In effects, it is often believed that what makes an action a good or bad action are the good or bad motives or intentions according to which the action is done. And, finally, how should a clear and precise analysis of our intuitions be done, if it is possible at all?

The question regarding the existence of moral intuitions is raised by persons that "deliberately deny that reflection enables them to discover any such phenomenon in their conscious experience as the judgment or apparent perception that an act is in itself right or good, in any other sense than that of being the right or fit means to the attainment of some ulterior end."[108] According to Sidgwick, this denial is opposed to common experience, and is probably a problem of understanding what is the subject of the analysis: persons who deny the existence of moral

---

(Sidgwick, 2000, p. xxviii), was so bored by Sidgwick's *Methods* that he never read another book on ethics. The dullness cannot be ascribed to an unsatisfactory capability of Sidgwick as a writer, but rather to the structure of the subject itself. Sidgwick's book is inspired by Aristotle's Ethics, and Aristotle is not less dry in his writing. Sidgwick may make things esthetically worse by adding to the difficulty of writing in a captivating manner this subject his rigor in researching concepts as definite as possible, but he is not entirely to be blamed. In (Parfit, 2011 a, p. xxxiv-xl) Parfit shows how Sidgwick is sometimes an excellent writer, but he cannot be engaging in the *Methods* due to clearness' sake.

[107] See chapter 2.2 for an explanation about Philosophical Intuitionism.

[108] (Sidgwick, 1874, p. 211)

intuitions confuse the matter of their *existence* with the matter of their *validity*.[109] The existence of a moral intuition implies merely that its truth appear immediate as soon as the intuition is present in our experience, and not after a reasoning of any kind. Intuitions can be incorrect, and can be corrected by reasoning and comparison with other intuitions, or with formulas that appears intuitively right from a moral point of view and that may help in accepting or discarding an intuition. A collection of general moral formulas, when shared by people of the same age and civilization and "regarded as a code imposed on an individual by the public opinion of the community to which he belongs",[110] is called by Sidgwick the Positive Morality of a community. The collection of general moral formulas that are conceived as approved by mankind,[111] becoming regarded as a "body of moral truth",[112] is called Morality of Common Sense.

Once separated the matters of existence and validity, an individual can understand if intuitions exist only by introspection. Many admit this existence.

Sidgwick believes that the second objection, according to which the focus of an analysis of moral intuitions should not be on the standard of virtues such as wisdom or bravery, but on the motives and intentions of the moral agent, is also partly due to a misunderstanding. Any moralist would agree that "moral judgments which we pass on actions relate primarily to intentional actions regarded as intentional":[113] it is not the act itself that we judge morally right or wrong, nor his consequences, but "the effects which [the agent] foresaw in willing the act; or, more strictly, his volition or choice of realising the effects as foreseen",[114] and those foreseen effects that an agent wants to produce are what is commonly called Intention. In other words, we judge right or wrong the intentions of an act. Stating

[109] Which may be linked by some author with the matter of their *origin*, and therefore someone may confound the three matters. The matter of the origin of moral intuitions may linked with their validity because "it has been often assumed, both by Intuitionists and their opponents, that if our moral faculty can he shown to be 'derived' or 'developed' out of other pre-existent elements of mind or consciousness, a reason is thereby given for distrusting it; while if, on the other hand, it can be shown to have existed in the human mind from its origin, its trustworthiness is thereby established." (Sidgwick, 1874, p. 212). According to Sidgwick, there is no ground for believing the truth of any of these two assumptions. For his full opinion on this subject, see (Sidgwick, 1874, p. 212-213)

[110] (Sidgwick, 1874, p. 215)

[111] "Or at least of that portion of mankind which combines adequate intellectual enlightenment with a serious concern for morality" (Sidgwick, 1874, p. 215)

[112] (Sidgwick, 1874, p. 215)

[113] (Sidgwick, 1874, p. 201)

[114] Ibid.

that an action is right or wrong because its intention is right or wrong is tautological, and analyzing the morality of common sense by means of the examination of intentions means nothing.

The case of motives is different. Whereas an intention is the volition or choice of realizing foreseen effect, a motive is the *desire* of some of the effects foreseen in an action.[115] Sidgwick cannot deny that common sense sometimes judges an action for a moral point of view examining its motives, but he believes that this examination is not the correct procedure for moral reasoning. In fact, it might be the case that the set of someone's foreseen consequences of an action, that is the intention of the agent, include wittingly, together with the desired consequences, that might be a good motive for performing the action, also foreseen consequences that are not desired and are seen as bad. No moralist would admit that someone is not responsible for those bad consequences simply because they are not desired.

Consider in fact two kind of unconventional cases, that are to be admitted since motives and intentions are different things: (a) that the intention of an act may be judged to be partially wrong while the motive is recognized as good, and (b) that the intention of an act may be judged partially right while the motive is recognized as bad. In those cases common sense seems perplexed in evaluating the morality of the action. Let us imagine some examples of those two kinds of cases.

Situation (1), kind of cases (a). An agent commits perjury to save a parent's life: the perjury is part of the intention of the agent, and is regarded as bad. The motive is saving a parent's life, and is regarded as good, but it is not clear to common sense if the good motive is enough for absolving the agent from the responsibility of committing perjury. Nevertheless, usually common sense admits actions like the perjury for saving someone, thus this situation might suggest that the rightness or wrongness of an action's motive is a sufficient and necessary condition for common sense for performing a moral judgement on the action.[116]

---

[115] It may appear that, according to common sense, motives and intentions are actually seen as synonymous terms. As I show in note 117, Sidgwick is right: sometimes the two terms are used as synonyms, but common sense clearly distinguishes the two concepts.

[116] Formally, if we call R(*) the rightness of an *, which can be an act, called A, an intention, called I, or a motive, called M, according to situation (1) might be believed that R(A)↔R(M). According to this formula, since in (1) R(M) is true, R(A) must be true.

Situation (2), kind of cases (a). A nihilist blows up a railway train containing an evil emperor and a great number of other innocent persons. The nihilist's motive is to kill the evil emperor, and it might be regarded as a good motive, but killing many innocent persons is a bad action.[117] In this situation, common sense is more perplexed to approve the action, and consequently appears less incline to admit that the morality of an action's motive is a sufficient condition for deciding if the action is right or wrong. In this case, it might be admitted that the life of the innocents must be considered as well, and therefore we must consider, when making moral evaluations, not the motives, but the intentions.[118]

Kind of cases (b) are not less complicated. Suppose a situation (3): a policeman prosecutes from malice a person whom he believes to be guilty.[119] If the policeman does his job repressing his malice, his act would be regarded as morally better. But it is generally admitted that we cannot always suppress entirely a strong emotion, and thus, though the policeman can refrain from acting only from malice, for example not causing the guilty person unnecessary pain, in this situation the policeman can act prompted by both his sense of duty and a wrong motive. Both motives lead to desires of the same consequences, and it seems impossible to perform an act aiming at satisfying only by the former desire and not the latter. The policeman would not be condemned by common sense if he acted prompted by both desires.

Imagine finally a situation (4), kind of cases (b). A policeman refuses to prosecute a person he knows to be guilty, because the policeman knows he would prosecute the person from malice, and believes that prosecute from malice is immoral. It seems difficult to approve this policeman's conduct: according to

[117] Notice that Sidgwick's distinction between intentions, that are volitions or choices of realizing foreseen effects, and motives, that are desires towards *some* foreseen effects, appears here clearly as accepted by common sense. In fact, it can be said that that the nihilist's intention is to kill the emperor, but no one would admit that *he did not intend* to kill other people. Intention are a set of foreseen effects, and in this set motives are included; thus is correct to say that the nihilist intends to kill the emperor. In the set is also to be included the consequence of killing others, therefore it is incorrect to state that the nihilist did not intend to kill them. On the opposite, it might be said that he had no motive to kill other people, since it was not his desire: his motive is merely killing the emperor.

[118] Formally, if we call R(*) the rightness of an *, which can be an act, called A, an intention, called I, or a motive, called M, according to situation (2) might be believed that $R(A) \leftrightarrow R(I)$. In fact, in this situation R(A) is believed true, but R(M) is believed false, thus $R(A) \leftrightarrow R(M)$ is false. But a formula such as $R(M) \rightarrow R(A)$ can be still believed to be correct.

[119] Sidgwick says in (Sidgwick, 1874, p. 202) that this case is suggested by Bentham.

common sense, it seems a duty to persecute a guilty person, at least for a policeman.[120]

Sidgwick concludes then "that while many actions are commonly judged to be made *better* or *worse* by the presence or absence of certain motives, our judgments of *right* and *wrong* strictly speaking relate to intentions, as distinguished from motives".[121]

> We may conclude then that the moral judgments which the present method attempts to systematise are primarily and for the most part intuitions of the rightness or goodness (or the reverse) of particular kinds of external effects of human volition, presumed to be intended by the agent, but considered independently of the agent's own view as to the rightness or wrongness of his intention; though the quality of motives, as distinct from intentions, has also to be taken into account.[122]

Those intuitions are liable of errors, that are resolved through comparison with other intuitions. Intuitions that can help correcting other intuitions can have the shape of general formulas, and are sometimes called principles, for example gratitude, temperance or law observance, consensually accepted by mankind and that are part of the morality of common sense.[123] Regarding those general principles, it must be noticed that, when we try to apply them, they seem to lack of clearness and precision.

> For instance, we should all agree in recognising Justice and Veracity as important virtues; and we shall probably all accept the general maxims, that 'we ought to give every man his own' and that 'we ought to speak the truth': but when we ask (1) whether primogeniture is just, or the disendowment of corporations, or the determination of the value of services by competition, or (2) whether and how far false statements may be allowed in speeches of advocates, or in religious ceremonials, or when made to enemies or robbers, or in defence of lawful secrets, we do not find that these or any other current maxims enable us to give clear and unhesitating decisions. And yet such particular questions are, after all, those to which we

[120] This is the counterexample of the formula R(M)→R(A). Here R(M) is true, since the motive of the act is the denial of malice, and malice is not admitted as a morally good motive. Malice would be a wrong motive: a ¬R(M). Acting from its denial, if classical logic is to be admitted at this basic level of moral reasoning, means to acting from ¬¬R(M), which is R(M). But R(A) is false, since it is wrong not to prosecute a guilty person: in this situation we have a ¬R(A). Thus, situation (4) can be represented as (R(M) ∧¬R(A)), which is the counterexample of R(M)→R(A).
[121] (Sidgwick, 1874, p. 204)
[122] (Sidgwick, 1874, p. 210)
[123] Intuitions can be corrected also by other kinds of formulas, established by philosophers, such as the Kantian formula or the so-called Golden Rule, that will be commented in chapter 4.4.

> naturally expect answers from the moralist. For we study Ethics, as Aristotle says, for the sake of Practice: and in practice we are concerned with particulars.[124]

Sidgwick tries therefore to define those formulae with enough precision to make them suitable as scientific axioms, in order to define exactly at least a part of the morality of common sense. But it is not clear if an examination of this kind should be done at all. In fact, it might be that moral rules are similar to the rules of art: rules and definite prescriptions may help in the performance, but "they can never do all (…) the highest excellence is always due to an instinct or tact that cannot be reduced to definite formulae."[125] In art, it is not possible to prescribe a certain method for producing beauty; it might be said that the same is valid for morality. This view is called by Sidgwick Aesthetic Intuitionism. According to Sidgwick, this view cannot help in cases of different moral opinions, and accepting it would endanger seriously the authority of ethics. Also, this view makes impossible decide on the validity of moral claims, since it would be possible only "examining in detail the propositions which have been put forward as ethical axioms". If some axiom is not clear enough to evaluate a principle, it must be searched a clearer axiom. It is probably true that precise and definite axioms cannot be found in adequate exactness merely by observing common moral reasoning of men, but it is also true

> that they are at least implied in these reasonings, and that when made explicit their truth is self-evident, and must be accepted at once by an intelligent and unbiassed mind. Just as some mathematical axioms are not and cannot be known to the multitude, as their certainty cannot be seen except by minds carefully prepared, but yet, when their terms are properly understood, the perception of their absolute truth is immediate and irresistible. Similarly, if we are not able to claim for a proposed moral axiom, in its precise form, an explicit and actual assent of "*orbis terrarum*," it may still be a truth which men before vaguely apprehended, and which they will now unhesitatingly admit.[126]

How, then, are those axioms to be found? What are its requirements? The conditions are four:

[124] (Sidgwick, 1874, p. 215)
[125] (Sidgwick, 1874, p. 228)
[126] (Sidgwick, 1874, p. 229)

1. "the terms of the proposition must be clear and precise",[127] as Bacon and Descartes prescribe, and for motives already indicated when opposing Aesthetic Intuitionism.[128]
2. "The self-evidence of the proposition must be ascertained by careful reflection",[129] which means that intuitions must not be confounded with impulses that has nothing to do with Reason, nor must be confounded with opinions, that may be common but unwarranted in their self-evidence.
3. "The propositions accepted as self-evident must be mutually consistent".[130]
4. The fact that no one disagrees regarding the truth of the axiom.

Nakano-Okuno notices[131] that Sidgwick applies, without stating it, also a fifth requirement:

5. The fact that the axiom must not be tautological.

In all moral principles analyzed by Sidgwick, those four conditions cannot be met unless recurring to an ultimate end. He will show that this ultimate end can only be General Happiness. I now present Sidgwick's examination of Justice as an example for all his analysis regarding moral principles of common sense.

It is worth to remember that this analysis is *not* an attempt to give a definition valid for every part of the usage of the term, "for many persons are undoubtedly vague and loose in their application of current moral notions".[132] Justice is believed by common sense to be a quality desirable to realize in the conduct and social relations of men; Sidgwick's aim is to find a clear, explicit formulation whose truth would appear as self-evident to everyone, and which is not in contradiction with other moral principles admitted as true.

---

[127] (Sidgwick, 1874, p. 338)
[128] Wittgenstein, or one of his readers, would disagree on the importance of a clear definition. For example, He might say that, about moral concepts, there is no need of precise definitions or boundaries: "We do not know the boundaries because none have been drawn. [. . .] Does it [. . .] make the concept usable? Not at all (Except for that special purpose.) No more than it took the definition: I pace = 75 cm. to make the measure of length 'one pace' usable. And if you want to say "But still, before that it wasn't an exact measure", then I reply: very well, it was an inexact one.— Though you still owe me a definition of exactness." (Wittgenstein, 1953). Wittgenstein might be right in some cases, but not in the case that Sidgwick's treatise, since sometimes moral intuitions seems to conflict with each other, and if we do not know the boundaries we do not know if they actually do conflict. We cannot omit to check if moral axioms are mutually consistent.
[129] (Sidgwick, 1874, p. 339)
[130] (Sidgwick, 1874, p. 341)
[131] (Nakano-Okuno, 2011, p. 87-89)
[132] (Sidgwick, 1874, p. 264)

3.2. Justice

According to Sidgwick, in the whole third book "there is no case where the difficulty is greater, or the result more disputed, than when we try to define Justice."[133] In fact, at first glance it might seem that a definition of justice may be found examining its relation with law. Unfortunately, we admit that some laws can be unjust. Also, we ascribe to justice behaviors that have nothing to do with law: for example, men of common sense would agree in stating that a father should be just with his children, but they would not refer to any legal obligation in this statement. Therefore, Justice must be distinguished from the duty of law observance.[134] In order to find a definition, we must comprehend observing what kind of law justice is realized according to common sense.

Sidgwick's answer is that "the laws in which Justice is or ought to be realised, are laws which distribute and allot to individuals either objects of desire, liberties and privileges, or burdens and restraints, or even pains as such".[135] This seems indeed a step forward, but how is this distribution to be realized?

Perhaps, equally: justice seems to have great affinity with equality. But, for example, give something more for people in a difficult situation it is not thought to be unjust, as it is not necessarily thought to be unjust, for example, that men, but not women, should enter the military and fight for their country. Clearly, none of the two cases has anything to do with equality. No doubt, if an act affects iniquitously the interest of individuals, and this inequality appears arbitrary, the act is not believed to be just. Still, equality is not a concept that can help understanding how ought to be realized a distribution that respects the criteria of Justice.

A more appealing guess is that a just distribution satisfies all claims that are recognized as valid, without advantaging or disadvantaging none of them who claim something. Now we need to understand what claims are to be recognized as valid. Those are: the one based on agreements between individuals or groups of individuals, the ones based on natural rights, and the ones based on desert. But all

[133] Ibid.
[134] Examined by Sidgwick in (Sidgwick, 1874, p. 295-311)
[135] (Sidgwick, 1874, p. 265-266) Pains are obviously reserved for people who break the law, and must be allotted by law itself.

three have problems.

The problem with agreements is that they may be thought as binding not only for what is explicitly stated by the two parts at the moment of making the agreement, but also of tacit understandings.

> But this latter term is a difficult one to keep precise: and, in fact, is often used to include not only the case where A has in some way positively implied a pledge to B, but also the case where B has certain expectations of which A is aware. [. . .] if the expectation was such as most persons would form under the circumstances, there seems to be some sort of moral obligation to fulfil it, if it does not conflict with other duties, though the obligation seems less definite and stringent than that arising out of contract.[136]

Even if there is an obligation to fulfil those expectations, this obligation is not clear. It is natural "to expect that any particular man will do as others do in similar circumstances, and, still more, that he will continue to do whatever he has hitherto been in the habit of doing [. . .] On the other hand, if a man has given no pledge to maintain a custom or habit, it seems hard that he should be bound by the unwarranted expectations of others."[137] The morality of common sense is perplexed about this case.

A kind of natural expectation that may be intended as the force that reforms the constantly varying customs, and that is the second kind of claim that seems to be always recognized as valid, is the respect of natural rights. In fact, there might be some sort of natural rights that every individual needs to see preserved, and thus people changes customs when they see that this change better protects natural rights.

Even though the notion of natural rights is not a notion definite and clear to common sense, some thinkers[138] have systematized this notion and brought all natural rights under one principle, which is freedom. But the notion of freedom has problems, too.

In fact, it must be admitted that freedom cannot be enjoyed by anyone: children, or insane people, will better off if coerced for their own good. But how far

[136] (Sidgwick, 1874, p. 269)
[137] (Sidgwick, 1874, p. 270)
[138] Sidgwick does not specify any precise philosopher. He is probably referring to Mill, frequently quoted in *The Methods of Ethics*, and maybe Von Humboldt. Sidgwick, for example, seems to implicitly refer to their reasoning concerning Freedom in (Sidgwick, 1874, p. 444).

should a sane adult be admitted to be free, if she can have a self-damaging behavior, or is not enough intelligent to provide for herself better than what other would provide for her, seems something upon which decision can be made only referring to general happiness. But it would mean admit an utilitarian basis of the principle of Justice or of the principle of Freedom, and that would lead us to wonder to what extent Justice and Freedom are actually principles that can be subjects of an intuition. Furthermore, the term "freedom" is ambiguous: if freedom means that anyone is free to do anything, freedom would "allow any amount of mutual annoyance except constraint".[139] If freedom, on the opposite, forbids mutual annoyance, the freedom of action would result limited to an inacceptable point. "Hence in distinguishing the mutual annoyances that ought to be allowed from those that must be prohibited we seem forced to balance the evils of constraint against pain and loss of a different kind":[140] once again, the Utilitarian solution. Also, it is dubious if freedom should allow mutual agreements that limit significantly freedom itself between individuals or groups of individuals. For example, it is not clear if a man has the freedom to sell himself as a slave. Utilitarianism seems to provide the only reasonable solution in this case too. But maybe freedom is not the ultimate end of distributive Justice: it seems that not only freedom, but also all other benefits and burdens should be distributed. In fact, Desert can be a criterion for distribution more effective than freedom.

But, as we said, Desert is problematic too: does someone deserves something as determined by her results?

> For it may be said that the actual utility of any service must depend much upon favourable circumstances and fortunate accidents, not due to any desert of the agent: or again, may be due to powers and skills which were connate, or have been developed by favourable conditions of life, or by good education, and why should we reward him for these ?[141]

It does not seems that someone deserves something according to results. On the opposite, is desert determined by how much was someone's effort? According to Determinists, an agent has no desert even for this aspects, hence probably free will

[139] (Sidgwick, 1874, p. 275)
[140] (Sidgwick, 1874, p. 275-276)
[141] (Sidgwick, 1874, p. 283)

should be admitted.[142] But not even admitting free will is a solution, "For in any case it does not seem possible to separate in practice that part of a man's achievement which is due strictly to his free choice from that part which is due to the original gift of nature and to favouring circumstances".[143] Therefore neither desert does not seem to provide acceptable answers.

A solution might to be the notion of fitness:

> We certainly think it reasonable that instruments should be given to those who can use them best, and functions allotted to those who are most competent to perform them: but these may not be those who have rendered most services in the past.[144] And again, we think it reasonable that particular material means of enjoyment should fall to the lot of those who are susceptible of the respective kinds of pleasure; as no one would think of allotting pictures to a blind man, or rare wines to one who had no taste.

We cannot appeal to this solution, because it is basically a Utilitarian solution.[145] The axiom of Justice at which Sidgwick aims cannot be found. Maybe it is impossible to find it if we do not agree in basing it on Utilitarian ground.

The research of all other axioms, that I will not summarize for briefness' sake, has the same result: Sidgwick is unable to find the aimed axioms unless he recurs to the utilitarian principle.

To summarize, we found out that Justice is the principle according to which the distribution of a certain good, material or non-material, satisfies all valid claims – that is to say claims justified by agreements, freedom or desert – but still there are some relevant unanswered questions.

---

[142] Sidgwick writes: "The only tenable Determinist interpretation of Desert is, in my opinion, the Utilitarian: according to which, when a man is said to deserve reward for any services to society, the meaning is that it is expedient to reward him, in order that he and others may be induced to render similar services by the expectation of similar rewards." (Sidgwick, 1874, note 1 to pg. 284)

[143] (Sidgwick, 1874, p. 285)

[144] And thus fitness conflicts with the imprecise conception of desert. Common sense seems to prefer, in those kind of situations, to act according with desert. Therefore fitness, that is actually an utilitarian solution, can be said to conflict with common sense in cases like these.

[145] Furthermore, we cannot appeal to this solution because it sometimes may diverge from what common folk perceives as justice, because it may conflict with other principles that are conceived as part of the notion of justice, for example with the concept of desert as highlighted by the previous note. Still, if *both* fitness and a conflicting principle, such as desert, are based on utilitarian ground, the conflict disappears.

- Are implicit claims valid in agreements?
- How far is just to pursue freedom, conceived as "perfect mutual non-interference of all the members of a community as an absolute end"?[146]
- How should a moral agent evaluate Desert? Should we reward effort or results?

Those problems might be solved by Utilitarianism, but before admitting that the morality of common sense has its implicit basis on Utilitarianism we should verify if it has not its basis in more general formulas, recognized trough reasoning by sages and philosophers.

3.3. Philosophical Intuitionism

A moral philosopher is thought to tell people what they ought morally to do, and not what they morally do, and thus, for a philosopher, it is possible to diverge to a certain extent from what is believed by Common Sense. But a philosopher's work will always be tested by its conclusion: if they are in flagrant conflict with commonly shared customs, a moral philosopher's work will not be approved, at least by his peers. The difficulty of being approved by his community challenged some philosophers not because they struggled to find agreement with common folk, but because, in attempting to find self-evident formulas, they happened to find formulas that, if analysed enough, are tautological.

For example Plato, in order to provide a scientific knowledge on ethical matters, prompted by Socrates' life and teaching, stated that the practical part of ethical science consist mainly in the knowledge of virtue and how to exercise it. This seems reasonable, but when we try to understand what Plato intended for virtue, we find that virtue is "the knowledge of what is good in certain circumstances and relations, [and in] a harmony of the different elements of man's appetitive nature, that their resultant impulse may be always in accordance with this knowledge. But it is just this knowledge (or at least its principles and method) that we are expecting

[146] (Sidgwick, 1874, p. 444)

him to give us".[147] Other philosophers have identified the knowledge of virtue as the basis of their ethical system: for example Aristotle,[148] or the Stoics.[149]

Despite the difficulties, in western tradition can be found a self-evident, non-tautological general formula that is remarkable for moral philosophers. It is the so-called Golden Rule, that is generally stated in a form similar to "Do to others as you would have them do to you". Sidgwick believes this statement to be effective, even if imprecise. It is imprecise firstly because someone might "wish for another's cooperation in sin, and be willing to reciprocate it".[150] Secondly, it seems that it is not always true that we ought to do to others only what we think it right for them to do to us, "for no one will deny that there may be differences in the circumstances and even in the natures of two individuals, A and B, which would make it wrong for A to treat B in the way in which it is right for B to treat A."[151]

According to Sidgwick, a more precise, self-evident, non-tautological formula of the Golden Rule can be found. He calls it

> *The principle of Equality*: it cannot be right for A to treat B in a manner in which it would be wrong for B to treat A, merely on the ground that they are two different individuals, and without there being any difference between the natures or circumstances of the two which can be stated as a reasonable ground for difference of treatment.[152]

---

[147] (Sidgwick, 1874, p.376)

[148] Which defined virtues differently from what Plato did, but "telling us that the Good in conduct is to be found somewhere between different kinds of Bad [. . .] at best only indicates the whereabouts of Virtue: it does not give us a method for finding it" (Sidgwick, 1874, p.376).

[149] Sidgwick shows in (Sidgwick, 1874, p. 376-378) that the Stoics incur also in the following logical circle: according to their system, virtue is to behave according to reason, because following reason is natural. When asked why we should follow nature, Stoics answer that we must do it because it is reasonable. In those pages Butler is shown by Sidgwick to make the same mistake, even if in a more subtle form than Stoics do.

[150] (Sidgwick, 1874, p.380)

[151] Ibid.

[152] (Sidgwick, 1874, p.380). According to Sidgwick, Kant's formal rule of acting on a maxim that one can will to be universal, if "duly restricted" is an "immediate practical corollary" of the principle of Equality. On the restriction, see (Sidgwick, 1874, p.317-319) on veracity, and (Sidgwick, 1874, p.485-492) for the fact that sometimes an act is moral even if it cannot be wanted for it to be imitated. Other two corollaries of the Golden Rule can be found. One is: "if a kind of conduct that is not right (or wrong) for me is not right (or wrong) for some one else, it must be on the ground of some difference between the two cases, other than the fact that I and he are different persons." (Sidgwick, 1874, p.379). The second can be found in the same page. Here Sidgwick writes: "A corresponding proposition may be stated with equal truth in respect of what ought to be done *to* - not *by* - different individuals." It must be noticed that, in Sidgwick, the latter two principles I presented in this note are not presented as a corollary of the principle of Equality, but as a different formulation of the Golden Rule. Nevertheless, I find the principle of Equality comprehensive of both formulas. He would

This principle has a recognized practical importance in common sense. Still, it does not provide complete guidance and merely "throw a definite *onus probandi* on the man who applies to another a treatment of which he would complain if applied to himself".[153] Nevertheless, it can be the fundament of other non-tautological self-evident truths.

In fact, we can find a "somewhat different application of the same fundamental principle that individuals in similar conditions should be treated similarly"[154] in the formulation of a maxim that seems to be the core of self-regarding duties, and can be viewed as the formula for Rational Self-love, or Prudence.[155] This formula prescribes that the moral agent should have "impartial concern for all parts of [her] conscious life",[156] or, in other words, "Hereafter as such is to be regarded neither less nor more than Now."[157] The fact that the maxim of rational self-love is a "somewhat different application of the same fundamental principle that individuals in similar conditions should be treated similarly" might have been a suggestion for Parfit's review of our concept of Personal Identity and for some of Nagel's ideas in *the Possibility of Altruism*, as will result clear in chapter 7 and 5 of the present work. Sidgwick adds further considerations, such as that "All that the principle affirms is that the mere difference of priority and posteriority in time is not a reasonable ground for having more regard to the consciousness of one moment that to that of another. The form in which it practically presents itself to most men is 'that a smaller present good is not to be preferred' to a greater future good (allowing for difference of certainty)"[158] and the fact that "the principle need not to be restricted to a hedonistic application: it is equally applicable to any other interpretation of 'one's own good' in which good is conceived as a mathematical

---

probably agree, since, whenever he summarizes the important principles, he presents the only formula of the principle of Equality.

[153] (Sidgwick, 1874, p.380)

[154] Ibid.

[155] Sidgwick treated Prudence from an Intuitionist's point of view in (Sidgwick, 1874, p.327-328)

[156] (Sidgwick, 1874, p.381)

[157] Ibid.

[158] Ibid.

whole, of which the integrant parts are realised in different parts or moments of a lifetime."[159]

Aiming to a maxim of Universal Benevolence corresponding to the maxim of Rational Prudence, Sidgwick obtains "the self-evident principle that the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other; unless, that is, there are special grounds for believing that more good is likely to be realised in the one case than in the other. And it is evident to me that as a rational being I am bound to aim at good generally, so far as it is attainable by my efforts, not merely at a particular part of it. "[160] Even if this principle seems self-evident, Sidgwick notices that it might be questioned from those who adheres to common sense, since it is commonly admitted that "practically each man, even with a view to universal Good, ought chiefly to concern himself with promoting the good of a limited number of human beings, and that generally in proportion to the closeness of their connection with him".[161]

It must be noticed that the fact that, if the belief that each man ought to benefit a limited number of human beings is true, then Utilitarianism should be discharged, because it does not distinguish between close people and other people, and therefore would be too demanding.[162] But let us focus on Intuitionism for the moment.

Let us restate more clearly the moral formulas admitted by Sidgwick as self-evident, non-tautological axioms of Intuitionism:

> *The principle of Equality*: it cannot be right for A to treat B in a manner in which it would be wrong for B to treat A, merely on the ground that they are two different individuals, and without there being any difference between the natures or circumstances of the two which can be stated as a reasonable ground for difference of treatment.

---

[159] Ibid. Even if it may not be clear what could be an "interpretation of 'one's own good' in which good is conceived as a mathematical whole" but the term "good" has nothing to do with happiness, this sentence is consistent to a development of Utilitarianism, that is called Consequentialism. Sidgwick's *Methods* seems indeed to have great responsibilities in prompting this development. Nevertheless, as it will be shown in the conclusion of the present summary on the *Method*'s book in Intuitionism, according to Sidgwick it does not exist an "interpretation of 'one's own good' in which good is conceived as a mathematical whole" but the term "good" has nothing to do with happiness.

[160] (Sidgwick, 1874, p.382)

[161] (Sidgwick, 1874, p.382)

[162] Parfit's reply on this matter will be summarized in note 273.

The principle of Prudence: the moral agent should have impartial concern for all parts of his or her conscious life.
Alternative formulation of the principle of Prudence: a smaller present good is not to be preferred to a greater future good.

The principle of Benevolence: the good of any one individual is of no more importance than the good of any other, unless there are special grounds for believing that more good is likely to be realized in the one case than in the other.

As previously stated, the principle of Equality does not give complete guidance. In addition, it must be noticed that the principles of Prudence and Benevolence lacks of a clear definition of what "good" is. If it has to be admitted that "good" is happiness, the principle of Prudence is equivalent to the fundamental axiom of Rational Egoism and the principle of Benevolence is equivalent to the fundamental axiom of Utilitarianism. Since common sense seems to regard Benevolence as superior to Prudence, at least as a fundamental moral principle, and since in Rational Egoism the principle of Equality does not seem to apply, Intuitionism would overlap to Utilitarianism. Do we have any alternative to the identification between the good according to Intuitionism and general happiness?

In the history of Ethics, Virtue has been often identified with ultimate good, but it has been showed at the beginning of 3.3 that this conclusion involves a logical circle. Another principle that has been identified with ultimate good is goodness of will, but common sense does not seem to identify the goodness of an action according to the degree of goodness of the motive of the agent, as it has been attempted to demonstrate on chapter 3.1. Finally, sometimes perfection has been pointed put as ultimate good, but reflection suggests that the excellent quality of perfect skills and gifts "are only valuable on account of the good or desirable conscious life in which they are or will be actualized".[163]

It seems impossible to find an ultimate good alternative to Happiness.[164] Still, Utilitarianism cannot be accepted yet as definite form of common sense morality. In fact, it might be true that common sense needs an ultimate good in order to define

[163] (Sidgwick, 1874, p.395)
[164] It must be noticed that not even Aesthetics Intuitionism can refrain from having a clear notion of ultimate good, since all its indefinite maxims contain, even if latently, reference to a good as ultimate aim. For Sidgiwck's whole argument on this subject see (Sidgwick, 1874, p.392).

clearly its principles and it might be true that it cannot be found other satisfying definition of good than happiness, but it is *not* true that common sense perceives general happiness as ultimate good of morality. Before accepting Utilitarianism as the foundation of Intuitionism, it is necessary to show why common sense does not perceive general happiness as ultimate good; if this is shown and it is demonstrated that common sense's value can actually be described precisely only in utilitarian terms, it would be possible to admit that common sense's acceptance of general happiness as ultimate good is real even if, so to say, unconscious.[165]

3.4. Happiness for Intuitionism (and Justice again)

Sidgwick finds that there are four motives responsible of common sense's refrain from accepting happiness, meant as surplus of pleasure over pain, as ultimate good.

1. No words can be clearer than Sidgwick's for describing the first motive:

> The term Pleasure is not commonly used so as to include clearly all kinds of consciousness which we desire to retain or reproduce: in ordinary usage it suggests too prominently the coarser and commoner kinds of such feelings [. . .]
>
> Also, our knowledge of human life continually suggests to us instances of pleasures which will inevitably involve as concomitant or consequent either a greater amount of pain or a loss of more important pleasures: and we naturally shrink from including even hypothetically in our conception of ultimate good these [. . .] pleasures.[166]

2. We experience many important pleasures only if we desire other things than pleasure, as it has been discussed in chapter 1.4.

3. When it is stated that happiness is the ultimate good of man, often it is understood as personal happiness, as for example in Hobbes' Leviathan. Common sense cannot accept this identification, but neither Utilitarianism can.

4. On the fourth motive it is again worth to quote fully Sidgwick:

> it seems true that Happiness is likely to be better attained if the extent to which we set ourselves consciously to aim at it be carefully restricted […]not only because action is likely

[165] Also Mill tried to show that actually common sense morality accepted general happiness as unconscious ultimate aim (Mill, 1952, pp. 6,7,16,17). According to Sidgwick, his demonstration is mistaken. See (Sidgwick, 1874, p.387-388).

[166] (Sidgwick, 1874, p.402). The issue according to which we experience pleasures at the cost of more important pleasures has been discussed in chapter 2.3 of this work.

> to be more effective if our effort is temporarily concentrated on the realisation of more limited ends [. . .] but also because the fullest development of happy life for each individual seems to require that he should have other external objects of interest besides the happiness of other conscious beings. And thus we may conclude that the pursuit of the ideal objects before mentioned, Virtue, Truth, Freedom, Beauty, etc. *for their own sakes*, is indirectly and secondarily, though not primarily and absolutely, rational; on account not only of the happiness that will result from their attainment, but also of that which springs from their disinterested pursuit.[167]

Still, there seems to be no alternative candidate than Happiness for being the greater good, as showed in 3.2 and 3.3. If we accept Utilitarianism as fundament for the morality of common sense, this latter can be described and analyzed in scientific axioms. In effects Sidgwick shows, in a long and detailed exposition,[168] how all principles of the morality of common sense are to be precisely described invoking general happiness as paramount aim:[169] this means that the morality of common sense overlaps with Utilitarianism and can be precisely defined only through it. In other words, "the Morality of Common Sense may be truly represented as at least unconsciously Utilitarian".[170] It must be noticed that Intuitionism and Utilitarianism have a coincidence that is only *general* and *qualitative*, since details may vary in the two methods. But there is no need to have complete coincidence:

> Utilitarians are rather called upon to show a natural transition from the Morality of Common Sense to Utilitarianism, somewhat like the transition in special branches of practice from trained instinct and empirical rules to the technical method that embodies and applies the conclusions of science: so that Utilitarianism may be presented as the scientifically complete and systematically reflective form of that regulation of conduct, which through the whole course of human history has always tended substantially in the same direction.[171]

---

[167] (Sidgwick, 1874, p.405-406)
[168] (Sidgwick, 1874, p.426-459)
[169] In (Sidgwick, 1874, p.423-426) Sidgwick points out that, in his connection of the maxims of Intuitionism to the fundamental axiom of Utilitarianism, he does an operation that has been previously successfully attempted by David Hume in (Hume, 1751). In this book, Hume proved that virtues are actually felicific features. It must be noticed that Sidgwick's analysis is certainly more detailed (even if probably less engaging) than Hume's. Furthermore, Hume was obviously less aware of the limits and the strengths of Utilitarianism, since the school itself of Utilitarianism begins after Hume's death. In fact, Hume dies in 1776. The Utilitarian school is considered starting with Bentham's *An Introduction to the Principles of Morals and Legislation*, published first in 1789.
[170] (Sidgwick, 1874, p.424)
[171] (Sidgwick, 1874, p. 425)

For our purpose, there is no need to summarize the whole part in which Sidgwick demonstrates the coincidence of every Intuitional principle with Utilitarianism.[172] I need only, for completeness' sake, to show how the theoretical problems regarding the principle of Justice are indeed solved by the Utilitarian theory.[173]

As seen before, Justice is the principle according to which the distribution of a certain good, material or non-material, satisfies all valid claims. Claims are recognised as valid if justified by agreements, freedom or desert. However, as we saw in 3.2, this formulation is unacceptable, since some points are too vague and there are some problems. In order to solve this, we asked ourselves what the theoretical prerequisite of the principle was.

If it is assumed an utilitarian perspective, our theoretical prerequisites are the Principle of Benevolence and the principle of Equality. Let us try to resolve, one by one, the difficulties found in chapter 3.2.

The first question was: are implicit claims valid in agreements? Agreements are considered binding for what is explicitly stated, but it is uncertain if they are binding for what is implicitly intended. According to the Utilitarian theory, disappointment of expectations is *pro tanto* an evil, and the greater the previous security of the expectant individual, the greater the evil: disappointed expectations undermine the trust in other people. Expectations should be disappointed only if the satisfaction of an expectation implies the sacrifice of a greater good, but according to Utilitarianism also explicit agreements have to be ignored if they cost the loss of a good greater than the one obtained through the agreement. In other words, according to the Utilitarian theory implicit and explicit expectations are to be treated equally. The fact that an explicit agreement can be broken if it costs the loss of a greater good does not conflict with common sense. For example, if it has been done an agreement that had to be realized later in time, and in this time an event changes the conditions so that the realization of the agreement would damage all persons involved in the agreement, it is not commonly seen as immoral to ignore the agreement.[174]

[172] Such an exposition can be found both in (Sidgwick, 1874, p.426-459) and in (Hume, 1751).
[173] The arguments here used in order to demonstrate how the problems of definition of Justice are solved by the Utilitarian theory are actually part of the fourth book of *the Methods of Ethics*.
[174] I recognize that this argument may not be convincing if exposed in such a brief manner. I recommend (Sidgwick, 1874, p. 443-444) for a more complete exposition and other arguments.

The second question was: how far is just to pursue freedom, conceived as perfect mutual non-interference of all the members of a community as an absolute end? The utilitarian motive for "leaving each rational adult free to seek happiness in his own way"[175] is that, as Von Humboldt[176] and Mill[177] suggest, usually

> each is best qualified to provide for his own interests, since even when he does not know best what they are and how to attain them, he is at any rate most keenly concerned for them: and again, the consciousness of freedom and concomitant responsibility increases the average effective activity of men: and besides, the discomfort of constraint is directly an evil and *pro tanto* to be avoided. [178]

Nevertheless, one's freedom may conflict with someone else's freedom: how much should freedom be limited? The Utilitarian axioms are clear: as far as it produces the greatest happiness. In effects, it is the only reason for which common sense imposes more constrains to the freedom of children and insane people than to sane adult people. Once again, there is no contrast between Utilitarianism and Intuitionism here. The former is simply more precise than the latter.

Another question was: how should a moral agent evaluate Desert? Should we reward effort or results? Clearly, Utilitarianism "encourages the production of general happiness by rewarding men for felicific conduct".[179] This means that "the Utilitarian scale of rewards will not be determined entirely by the magnitude of the services performed, but partly also by the difficulty of inducing men to perform them".[180] We saw that common sense does not have a principle for understanding if we should reward an action by its efforts or results; the answer might be that should be rewarded both, but effort should be privileged, since an encouragement is effective if it can motivate despite difficulties and efforts. A motivation toward a greater result can be conceived as a motivation toward a greater effort, because, if result is conceived as independent from effort, it would be trivial to encourage it. This is, once again, consistent with Intuitionism.

---

[175] (Sidgwick, 1874, p. 444)
[176] See his *the Limits of State Action*, first printed in 1792.
[177] (Mill, 1858)
[178] (Sidgwick, 1874, p. 444-445)
[179] (Sidgwick, 1874, p. 446)
[180] Ibid.

Things seem different in the case of punishment: according to Utilitarianism, punishment should be preventive, which means that it should discourage someone to commit actions that decrease general happiness. On the other hand, Intuitionism prescribes a punishment somehow correspondent to the damage done by who has to be punished. But, according to Sidgwick,

> We find that in the actual administration of criminal justice. Common Sense is forced, however reluctantly, into practical agreement with Utilitarianism. Thus after a civil war it demands the execution of the most purely patriotic rebels; and after a railway accident it clamours for the severe punishment of unintentional neglects, which, except for their consequences, would have been regarded as very venial.[181]

Here Sidgwick seems too optimistic. Even if his example can be accepted as a proof of his point, there are many counterexamples to the theory that, in practice, a punishment on Intuitional basis is equivalent to a punishment on Utilitarian basis. According to Utilitarianism, an institution such as jail would not be used as frequently as it is commonly done, because the damage inflicted by prison to the detained exceeds what is necessary as a deterrent for crimes. In other parts of the *Methods*, such as when discussing the concept of Fitness,[182] which I briefly sketched in 3.2, the author seems far more aware of this problem. As we saw in this section, this is not an argument against the consistency between Utilitarianism and Intuitionism, since the coincidence has not to be perfect. Nevertheless, this is a discrepancy between Utilitarianism and contemporary legal justice. An Utilitarian thinker may wonder if she ought to do something for changing her system of legal justice. Some answers to this question will be provided soon when analysing the book on Utilitarianism.

For now, since it has been showed

> how Utilitarianism sustains the general validity of the current moral judgments, and thus supplements the defects which reflection finds in the intuitive recognition of their stringency; and at the same time affords a principle of synthesis, and a method for binding the

[181] (Sidgwick, 1874, p. 446-447)
[182] He discusses it in (Sidgwick, 1874, p. 282-283)

> unconnected and occasionally conflicting principles of common moral reasoning into a complete and harmonious system,[183]

it is time to examine the last book of Sidgwick's *Methods*, entitled *Utilitarianism*.

Chapter 4: Utilitarianism

So far Utilitarianism has been described as the theory according to which a moral agent ought to produce the greatest general happiness. It is time to better understand what this means. In the book named after it Sidgwick tries to define better the term and explores the relationship between the three Methods. Since the incompatibility between Egoism and the other two Methods has been discussed in chapter 2 and since we already saw in the previous chapter the correspondence between Utilitarianism and Intuitionism, in the present work the exposition will be slightly different. I will summarize Sidgwick's try to better define Utilitarianism and I will examine only the discrepancies between common sense morality and Utilitarianism, asking how should an Utilitarian agent act when confronting those discrepancies. I will then conclude by summarizing the difficulties of the Utilitarian theory underlined by Sidgwick or arose while reading Sidgwick's work.

Before beginning the examination, it is important to notice that, according to Sidgwick, in the history of Ethics, the only Method that required a proof of its actual usage and its effectiveness was Utilitarianism. The first attempt to prove Utilitarianism was (Hume, 1751), another attempt is (Mill, 1861, p. 6,7,16,17), and a further attempt is (Sidgwick, 1874) itself. Sidgwick's book proofs successfully that Utilitarianism is commonly used for resolving the ambiguities of Intuitionism, but its success in proofing the effectiveness of Utilitarianism is only partially, even if impressively, achieved. Sidgwick says that "if Utilitarianism is to be proved to a man who already holds some other moral principles [. . .] it would seem that the process must be one which establishes a conclusion actually *superior* in validity to the premises from which it starts".[184] Since Utilitarianism solves some problems of Intuitionism, it might be thought as superior to Intuitionism. Unfortunately, as already pointed out, Sidgwick did not manage to show that Utilitarianism is also

[183] (Sidgwick, 1874, p.422)
[184] (Sidgwick, 1874, p. 419)

superior to Rational Egoism, leaving the Profoundest Problem unanswered and his proof half realized.

According to Sidgwick's historical reconstruction, contrary to Utilitarianism, in the history of Ethics there is little doubt that Egoism and Intuitionism are Methods actually used by people, and few suspected that they might be incoherent. Sidgwick's examination showed that Intuitionism is not, in fact, coherent if not guided by Utilitarianism. It might be asked if also Egoism is coherent or not. For a moment, Sidgwick dares to question the coherence of Egoism, but he does not develop his argument. His words are seminal for Nagel's and Parfit's work, and will offer the basis for solving the Dualism of the Practical Reason. He writes:

> From the point of view [. . .] of abstract philosophy, I do not see why the Egoistic principle should pass unchallenged any more than the Universalistic [that is another name for the Utilitarian principle]. I do not see why the axiom of Prudence[185] should not be questioned, when it conflicts with present inclination, on a ground similar to that on which Egoists refuse to admit the axiom of Rational Benevolence.[186] If the Utilitarian has to answer the question, 'Why should I sacrifice my own happiness for the greater happiness of another?' it must surely be admissible to ask the Egoist, 'Why should I sacrifice a present pleasure for a greater one in the future? Why should I concern myself about my own future feelings any more than about the feelings of other persons?'[187]

Sidgwick unknowingly ends his questioning with a *grand finale*, unaware of the weight that his words will have had in the successive century of ethical thinking:

> Grant that the Ego is merely a system of coherent phenomena, that the permanent identical 'I' is not a fact but a fiction, as Hume and his followers maintain; why, then, should one part of the series of feelings into which the Ego is resolved be concerned with another part of the same series, any more than with any other series?[188]

But Sidgwick did not dare further, and proceeds with different arguments, never writing anymore about this matter. The importance of this suggestion by Sidgwick

[185] The reader certainly remembers that the axiom of Prudence is the basis of Rational Egoism, as stated in 3.3
[186] The axiom of Rational Benevolence is the basis of Utilitarianism, as stated in 3.3
[187] (Sidgwick, 1874, p. 418)
[188] (Sidgwick, 1874, p. 419)

will be clear when analyzing part 3 of *Reasons and Persons*. Parfit writes that Sidgwick "went astray",[189] and shows why he went astray. We will briefly summarize Parfit's analysis in due time. It is now time to follow Sidgwick astray and focus on Utilitarianism.

4.1.The meaning of Utilitarianism: total and average principle

In order to better understand Utilitarianism, it might help being clear first on what Utilitarianism is not.

Utilitarianism is *not* "necessarily connected with the psychological theory that the moral sentiments are derived, by "association of ideas" or otherwise, from experiences of the non-moral pleasures and pains resulting to the agent or to others from different kinds of conduct".[190] This psychological theory is sometimes called Utilitarian theory of the origin of the moral sentiments. Sidgwick shows in (Sidgwick, 1874, p. 412) that this psychological theory might be accepted also by a Rational Egoist or an Intuitionist, and thus there is no necessary connection between this theory and the Method we are analyzing.

Utilitarianism does *not* imply that the *motive* of a moral action must always be the production of the greatest general happiness. In fact, "it is not necessary that the end which gives the criterion of rightness should always be the end at which we consciously aim".[191] The motive is not crucial, the only critical matter is that general happiness is obtained, "and if experience shows that the general happiness will be more satisfactorily attained if men frequently act from other motives than pure universal philanthropy, it is obvious that these other motives are reasonably to be preferred on Utilitarian principles".[192]

We can now examine closer the principle of Utilitarianism. It prescribes the greatest amount of general happiness. On happiness, and on the problems about evaluating and balancing pleasure and pain, it will suffice what previously stated in 1.4. With the term "general" it is intended, obviously, that everyone's happiness has to be considered. It is important to ask who "everyone" is: all *human* beings or all

[189] (Parfit, 1984, p. 141)
[190] (Sidgwick, 1874, p. 412)
[191] (Sidgwick, 1874, p. 413)
[192] Ibid.

sentient beings? As Sidgwick writes, "it seems arbitrary and unreasonable to exclude from the end, as so conceived, any pleasure of any sentient being".[193] Considering the happiness of all sentient beings certainly increases the difficulty in estimating pleasure and pain underlined in 1.4, since it is certainly more difficult to understand what are pain and pleasure for a mockingbird or for a poplar than understanding what are pain and pleasure for a man or a woman. Despite that, it would be paradoxical to think that non-human sentient beings have to be ignored when acting morally, and that their pleasures and pains have to be disregarded.

"But even if we limit our attention to human beings, the extent of the subjects of happiness is not yet quite determinate."[194] The problem of the indeterminateness of "the extent of the subject of happiness" will lead, approximately a century later than Sidgwick's book, to a branch of moral philosophy called Population Ethics, that concerns what size of human population should be the better size from a moral point of view. *Reasons and Persons* is seminal for Population Ethics, and Sidgwick's considerations on "the extent of the subject of happiness" certainly influenced it. Let us examine those considerations.

> In the first place, it may be asked, How far we are to consider the interests of posterity when they seem to conflict with those of existing human beings?[195]

The problem is not actually very challenging for the Utilitarian theory. Sidgwick writes:

> It seems [. . .] clear that the time at which a man exists cannot affect the value of his happiness from a universal point of view; and that the interests of posterity must concern a Utilitarian as much as those of his contemporaries, except in so far as the effect of his actions on posterity and even the existence of human beings to be affected must necessarily be more uncertain.[196]

Things get way more complicated if it is considered that "we can to some extent influence the number of future human (or sentient) beings. We have to ask how, on

[193] (Sidgwick, 1874, p. 414)
[194] Ibid.
[195] Ibid.
[196] Ibid.

Utilitarian principles, this influence is to be exercised."[197]

Sidgwick assumes "that for human beings generally, life on the average yields a positive balance of pleasure over pain."[198] He is aware that his assumption is controversial. Several thinkers would disagree: a famous example is Arthur Schopenhauer, whose *The World as Will and Representation* contains a chapter with the eloquent title *On the Vanity and Suffering of Life.*[199] Among recent thinkers, Christoph Fehige bases his view of Population Ethics on a principle[200] that implies that "Our world is worse than an empty world",[201] where "empty" means "not populated by any sentient being". Further authors might disagree with Sidgwick's assumption. Sidgwick believes that denying his assumption is

> clearly opposed to the common experience of mankind, as expressed in their commonly accepted principles of action. The great majority of men, in the great majority of conditions under which human life is lived, certainly act as if death were one of the worst of evils, for themselves and for those whom they love: and the administration of criminal justice proceeds on a similar assumption.[202]

If happiness is usually enough present in life to make it worth living, it might be thought that the greatest amount of general happiness can be obtained with the greatest number of people enjoying happiness. But it must be noticed that, if the resources are limited, an enhancement of the number of people might decrease the average happiness enjoyed by individuals. This raises another problem: what do we mean with "greatest amount"? Does Utilitarianism aims to the greatest *average* amount or to the greater *total* amount of happiness?

It is worth stating the principles clearly. I will use Parfit's words in doing that:

---

[197] Ibid.
[198] Ibid.
[199] It can be found on the second volume, as a supplement to the fourth book of the first volume.
[200] Namely, the Principle of Pareto-Superiority among Wishes, or POPSAW. For information about POPSAW read (Fehige, 1998).
[201] (Fehige, 1998, p. 521)
[202] (Sidgwick, 1874, p. 414-415) Sidgwick adds in a note at page 415 that "Those who held the opposite opinion appear generally to assume that the appetites and desires which are the mainspring of ordinary human action are in themselves painful". He demonstrates that it is not true in (Sidgwick, 1874, p. 45-47, note p.54-56). I omit a summary of this part because his argument is complex and not important for our purpose.

> *Average Principle*: If other things are equal, the best outcome is the one in which people's lives go, on average, best.[203]

More formally, if we call *N* the number of existing people, individual 1, individual 2,individual 3… individual n are called $p_1,p_2,p_3...p_n$, and $H(p_x)$ is a function which takes individuals as arguments and gives their happiness as values, the *Average Principle* would prescribe that the best moral choice is the one that maximize the result of the following formula:

$$(H(p_1) + H(p_2) + H(p_3) \dots + H(p_n))/N$$

Sidgwick,[204] as Parfit,[205] notices the success of this interpretation of Utilitarianism in economy.

> *Total Principle*: If other things are equal, the best outcome is the one in which there would be the greatest quantity of good.[206]

This time the formula that express the quantity to maximize is

$$(H(p_1) + H(p_2) + H(p_3) \dots + H(p_n))$$

Let us call this formula *U*. The Total Principle aims at maximize *U*, whereas the Average Principle aims at maximize *U*/*N*.

The contrast between the Average and the Total principle will dominate many pages of the fourth part of *Reasons and Persons*, as well as many pages of the third part of this work. Coming back to Sidgwick, we can say that he tried to solve a problem of Population Ethics nearly a century before this branch of applied ethics

---

[203] (Parfit, 1984, p. 386). The complete name should be "*the Impersonal Average Principle*". It will be clear why in 8.1.
[204] (Sidgwick, 1874, p. 415). His example is Malthus.
[205] (Parfit, 1984, p. 386). His example is Samuelson.
[206] In Parfit's formulation, (Parfit, 1984, p. 387) the word "good" is replaced by the expression "whatever makes life worth living", a notion that I choose not to define precisely in this work because it would require too much space. As for the previous principle, the complete name should be "*the Impersonal Total Principle*". It will be clear why in 8.1.

arose, and he understood that Utilitarianism is not a definite theory until this problem is not resolved. To my knowledge, he is the very first in doing that.[207]

What has been understood by theorists of Population Ethics but still not by Sidgwick is that both principles have consequences that are, at least at first glance, inacceptable. I will return soon on this concept. For now, let us return briefly on Sidgwick's book. If a total principle ought to be accepted, the losses of a part of the population can be balanced both by gains in happiness of another part of a population and by *additional happy people*. In Population Ethics, additional people are people that are caused to exist by the choice of a policy rather than another. For example, the choice of a family to have or not to have a child adds or does not add people. If happy people are added, the total sum of happiness increases, therefore a Total Principle would encourage this addition of happy people, and would approve the losses of a part of the population if balanced by new happy people.

Sidgwick accepts the total principle without giving reasons for his decision: simply, if the Utilitarian principle prescribes as ultimate aim happiness *on the whole*, it seems unreasonable to consider average happiness of *individuals*; thus the Average Principle should be discarded. Sidgwick, is aware that the conclusion according to which we must consider only the total sum of happiness and not the average happiness is in conflict with the view of common sense,

> Because its show of exactness is grotesquely incongruous with our consciousness of the inevitable inexactness of all such calculations in actual practice. But, that our practical Utilitarian reasonings must necessarily be rough, is no reason for not making them as accurate as the case admits; and we shall be more likely to succeed in this if we keep before our mind as distinctly as possible the strict type of the calculation that we should have to make, if all the relevant considerations could be estimated with mathematical precision.[208]

As previously stated, Parfit shows that the *Total Principle* conflicts with common sense also for more relevant reasons that cannot be so easily ignored. In fact, accepting that adding people enhanced the total quantity of good implies

[207] He is aware to be, at least, among the firsts to pose this kind of problems. He writes that this point "has not only never been formally noticed, but [. . .] seems to have been substantially overlooked by many Utilitarians." (Sidgwick, 1874, p. 415)
[208] (Sidgwick, 1874, p. 416)

> *The Repugnant Conclusion*: For any possible population of [. . .] people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living.[209]

The name that Parfit gives to this conclusion is a clear signal that he finds it unacceptable.[210] If *The Repugnant Conclusion* is unacceptable then the *Total Principle*, that implies *the Repugnant Conclusion*, should not be accepted too.

Unfortunately, the *Average Principle* is an alternative that leads to even less acceptable conclusions. Suppose for example that we live in a world with a population of 8 billion people, roughly as much as today, where everyone lives a really good life. Imagine that we are obliged to add people, and we have an alternative: we can choose (1) a policy that causes the existence of 4 billion people whose life will be good, but less good than the life of the already existing 8 billion people, whose lives are really good lives, or we can choose (2) a policy that causes the existence of a single person, whose life will be not worth living, and full of suffering rather than happiness. According to the *Average Principle*, we ought to choose (2), preferring the existence of a suffering person than the existence of happy people.[211] This does not seem more acceptable than *the Repugnant Conclusion*.[212]

---

[209] (Parfit, 1984, p. 388) . This can be demonstrated as follows. We call N the number of existing people, we call individual 1, individual 2,individual 3… individual n with the symbols $p_1,p_2,p_3...p_n$ and happiness enjoyed by an individual x is called $H(p_x)$. Remember that $U=(H(p_1) + H(p_2) + H(p_3) ... + H(p^n))$. Remember also that the *Total Principle* aims at maximize *U*. Let us assume that the *quanta* of happiness are measurable, and that a person who lives a life barely worth living enjoys only 1 *quantum* of happiness. Imagine any possible population of any value *N* and in which everyone lives a life that is more than barely worth living, and thus the value of *U*/*N* (average happiness) is greater than 1. Imagine now a possible population of *N'* people, where the number of people *N'* is greater than the *U* of the other population. Imagine now that everyone in this population of *N'* people has a life that is barely worth living (that has value 1). Since each person has only a quantum of happiness, the value of *N'* is equal to the value of the total amount of happiness *U'* of that population. By hypothesis, $N'>U$, but since $N'=U'$ we can conclude that $U'>U$. Therefore, the population in which everyone has a life barely worth living results better, *quod erat demonstrandum*.

[210] The unacceptability of *the Repugnant Conclusion* is debated. Some thinkers of Population Ethics, such as Michael Huemer (see (Huemer, 2008)), Törbjörn Tännsjö (see (Tännsjö, 2004)) and Yew-Kwang Ng (see (Ng, 1989)) believe that *the Repugnant Conclusion* ought to be accepted.

[211] I demonstrate it here. Suppose that the 8 billion people, that we call group A, have happiness 4 *pro capite*, the 4 billion people, or group B, have happiness 2 *pro capite* and the suffering person, or group C has happiness -8. The average happiness of (1) would be the average of groups A and B. The total happiness in A is $(8 \times 10^9) \times 4 = 32 \times 10^9$, the total happiness in B is $(-4 \times 10^9) \times 2 = 8 \times 10^9$, the average is the sum of the two total happiness divided by 12 billion people, that is the total of persons living in group A and B: the average happiness of (1) is thus $(32 \times 10^9 + 8 \times 10^9) / (12 \times 10^9) = 3,(3)$. On the other hand, the average happiness of (2) would be the average of the groups A and C. The total happiness in A is always $(8 \times 10^9) \times 4 = 32 \times 10^9$, the total happiness in C is $1 \times -8 = -8$, the average is

This means that Utilitarianism has not a clear principle for choosing a moral action that concerns future beings when the choice influences the number of people that will exist. The most intuitive principles, the *Average* and the *Total Principle*, seems opposed to common sense.

But there is another problem of Utilitarianism that Sidgwick is close to notice, and that later[213] will be pointed out by Robert Nozick. Sidgwick highlights

> that there may be many different ways of distributing the same quantum of happiness among the same number of persons; in order, therefore, that the Utilitarian criterion of right conduct may be as complete as possible, we ought to know which of these ways is to be preferred.[214]

Unfortunately, the Utilitarian formula does not give any guidance on "whether any mode of distributing a given quantum of happiness is better than any other."[215] On this matter, the principle that has been traditionally adopted by utilitarian thinkers previous to Sidgwick is Bentham's principle according to which "everybody to count for one, and nobody for more than one."[216] According to Sidgwick, this formula does not need any justification, since it seems a corollary of *the principle of Equality* and this *principle* is a self-evident moral truth.

But Bentham's principle according to which "everybody to count for one, and nobody for more than one" is *not* necessarily a corollary of *the principle of Equality*. *The principle of Equality* admits exceptions when it is true that more good is likely to be realized if not everyone count for one. Robert Nozick points out this aspect with those words:

---

the sum of the two total happiness divided by 8 billion people and one: the average happiness of (2) is therefore (32 x 10^9 + (-8)) / (8x10^9 +1)= 4 circa. The -8 of population C is trivial when counting the average. Since 4 is bigger than 3,(3), according to the *Average Principle* (2) is to be preferred to (1).

[212] Gustaf Arrhenius would say that the *Average Principle* violates the *Non-Sadism Condition*, a very intuitive principle according to which "an addition of any number of people with positive welfare is at least as good as an addition of any number of people with negative welfare, other things being equal" (Arrhenius, 2000, p. 203).

[213] In (Nozick, 1974)

[214] (Sidgwick, 1874, p. 416)

[215] Ibid. To be clear, it is worth to report the note on (Sidgwick, 1874, p. 417): "It should be observed that the question here is as to the distribution of *Happiness*, not the *means of happiness*. If more happiness on the whole is produced by giving the same means of happiness to B rather than to A, it is an obvious and incontrovertible deduction from the Utilitarian principle that it ought to be given to B, whatever inequality in the distribution of the *means* of happiness this may involve."

[216] (Sidgwick, 1874, p. 417)

> Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater gains in utility from any sacrifice of others than these others lose. For, unacceptably, the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility.[217]

If such an Utility Monster exists, it is not true that everybody counts for one, since sacrifices for the monster increase happiness: both the *Average* and the *Total Principle* would encourage the sacrifice of the total of happiness for the monster.[218] This seems to conflict with common sense to an extent so great that any Utilitarian thinker has to try to cope with the case of the Utility Monster. If an Utilitarian thinker does not, she should modify the theory of Utilitarianism or reject Utilitarianism altogether. The case of the Utility Monster is simply unacceptable for common sense, and this issue cannot be ignored by utilitarian thinkers.

Sidgwick succeeds in defining Utilitarianism more clearly than his predecessors, but his assumption on the *Total Principle* is controversial and Nozick's objection cannot be ignored. In the third part of this work I will try to cope with both. It is now time to focus on the discrepancies between Utilitarian and Intuitional theories, and on how should an Utilitarian behave when she meets them.

4.2. Utilitarian limits of conduct and problems

Since the Intuitional Method is Utilitarian only "unconsciously", it rejects happiness as ultimate end, for reasons explained in 3.4. What, then, is the basis of the Intuitional Method in its not reflected application? Under the influence of Adam Smith's *Theory of Moral Sentiments,*[219] Sidgwick writes that the moral sentiment is

---

[217] (Nozick, 1974, p. 41)

[218] It will be demonstrated in this note. Let us decide, by hypothesis, that each *quantum* of happiness sacrificed to the monster by a common individual result in tenfold happiness for the Monster. Let us assume, as Sidwick does, that lives are actually, on average, happy, and thus the total sum of pleasure minus pain gives a positive value different from 0. Let us call *m* the monster, which is not an individual that counts as one, and hold all other formal definitions previously given with reference to the discussion of *Total* and *Average Principles*. The sacrifice of everything by $p_1,p_2,p_3...p_n$ would result in a loss of happiness equal to U, decreasing the total by U. The gain of the monster if the sacrifice is made is equal to $10U$, thus $H(m)=+10U$. Therefore, if the sacrifice is made, the total would be $-U + H(m) = -U + 10U = 9U$. Allow the hypothesis that the value of $H(m)$ is equal to 0 if the sacrifice is not made, in order not to add or subtract any value to the formula if the monster receives nothing. In both formulas, Utilitarianism recommends to make the sacrifice, since $U/N < 9U/N$ and $U < 9U$.

[219] In a note Sidgwick supports Smith's view also with a brief analysis of penal codes of primitive communities (Sidgwick, 1874, p. 463).

mainly prompted not by Utilitarian considerations, but by sympathy, or "the echo [. . .] of each agent's passion in the breast of unconcerned spectators".[220] In Intuitionism, two kinds of sympathy determine moral sentiments: the sympathy concerning approval and disapproval of actions and sympathy "with the effect of conduct on others".[221] According to Sidgwick, the principles of Intuitionism arose from a compromise between those two kinds of sympathy. Experiences of pleasure and pain influence sympathy, but do not determinate it, and that is why Utilitarianism and Intuitionism sometimes diverge. The less the "accuracy with which the whole sum of pleasurable and painful consequences, resulting from any course of action, has been represented in the consciousness of an average member of the community",[222] the wider is the discrepancy. Many might be the causes of imprecision of this representation of the sum of pleasurable and painful consequences: inadequate knowledge of natural causes, habits of obedience to strong authorities, influence of "false religion",[223] the fact that "the sympathy of an average man with other sentient beings [. . .] has been much more limited than the influence of his actions on the feelings of others"[224] and others.

It might seem that, since the rules of common sense morality are not perfectly felicific, and thus not perfectly moral, a Utilitarian thinker ought morally to change common sense morality in order to make it more felicific. The objection that, even if Utilitarianism is a Method more definite than Intuitionism, it has still some uncertainties, such as the difficulty in hedonistic comparison, might not suffice to state that Utilitarian thinkers have no responsibility to reform the morality of common sense.

Sidgwick seems scared by the possibility of this reform, for several reasons. One reason is that, if an individual has a moral code, she is not to be considered as someone "for whom a code is yet to be constructed *de novo*".[225] Moral habits and tastes of existing people are part of their nature, and can only be partially modified by reasoning.

---

[220] (Sidgwick, 1874, p. 463)
[221] Ibid.
[222] (Sidgwick, 1874, p. 464)
[223] Ibid.
[224] Ibid.
[225] (Sidgwick, 1874, p. 468)

According to Sidgwick, it might not be possible even to approximate the current moral rules to Utilitarianism, since "since any particular existing moral rule, though not the ideally best even for such beings as existing men under the existing circumstances, may yet be the best that they can be got to obey",[226] and forcing men would be more harmful than felicific.

Furthermore, "the endeavour gradually to approximate to a morality constructed on the supposition that the non-moral part of existing human nature remains unchanged, may lead us wrong":[227] mankind's knowledge, sympathies, impulses vary constantly and changes in the corresponding established moral rules are required *immediately*, "in order that the greatest possible happiness may be attained *by the human being whose life is thus modified*":[228] changing morality both approximating Utilitarianism and not overlooking the changing non-moral part of human nature is risky.

If it might be better to refrain from a radical reform of common sense morality, it seems reasonable that the Utilitarian diverges sometimes from what the Intuitional Method would recommend.

"Generally speaking, he will clearly conform to it, and endeavour to promote its development in others. For, though the imperfection that we find in [. . .] Morality [. . .], we are much less concerned with correcting and improving than we are with realising and enforcing it."[229] Nevertheless, the Utilitarian must aid improving the existing moral order, and aid can be given only slightly diverging from it; the Utilitarian must consider carefully the consequences of this divergence. She must compare

> the total amounts of pleasure and pain that may be expected to result respectively from maintaining any given rule as at present established, and from endeavouring to introduce that which is proposed in its stead. That this comparison must generally be of a rough and uncertain kind, we have already seen; and it is highly important to bear this in mind; but yet we seem unable to find any substitute for it.[230]

[226] (Sidgwick, 1874, p. 469)
[227] Ibid.
[228] Ibid. Italic is mine.
[229] (Sidgwick, 1874, p. 475)
[230] (Sidgwick, 1874, p. 477)

Sidgwick suggest the general lines that an Utilitarian thinker must be careful not to overstep, in order to avoid to cause more damage than happiness. He first distinguishes between duty, that is blameworthy if disregarded but not necessarily praiseworthy if observed, and excellent conduct, that is not conceived as mandatory, but if followed is praiseworthy.[231] No act that an Utilitarian will do concerning action classified as excellent conduct seem to conflict with common sense, so he should not be aware of any particular warning. We will thus focus on the conflicts between Utilitarian rules and what is commonly conceived as duty.

When promoting a conduct more felicific, but contradicting duty, it must be considered the social disapprobation that will be met by the agent, and "its indirect effect in diminishing his power of serving society and promoting the general happiness in other ways."[232] Some rules, however felicific, are not worth this price: a new Utilitarian rule might be ignored by the community in which is proposed, because too demanding, or too difficult to comprehend, or too complex to apply.

Furthermore, the mere violation of a recognized moral rule may reduce the trust of fellow men in their moral rules, pushing them slightly towards moral anarchy. A similar effect might be present also in the agent, that might feel herself detached to some extent from customary morality, and this kind of morality, built on habits and sentiments, might be more trustworthy than a morality built on reasoning under certain conditions, such as when the moral decisions has to be done with little time and much pressure. Also, the effect of sympathy, that encourages the agent to perform a moral action, may decrease.

On the other hand, if it is possible to add a new Utilitarian rule that does not conflict with the rules of common sense, and if the rule is indeed felicific, there seems to be no danger in promoting the rule, and an Utilitarian thinker ought to promote it. This situation is not very common, but it can happen. A difficulty arises if, by imposing a rule, there is a condemnation of all who are not prepared to adopt it, which leads to pain for who these latter people and to a weakened effect of the moral example of the agent, since she will have shown an aggressive conduct. In this case,

[231] This distinction has been more carefully examined in (Sidgwick, 1874, p. 217-230). Sidgwick notices (Sidgwick, 1874, p. 492) that "this distinction between Excellence and Strict Duty does not seem properly admissible in Utilitarianism". Parfit would disagree, stating that the distinction between Excellence and Strict Duty exists also in Utilitarianism: see note 273 on this matter.
[232] (Sidgwick, 1874, p. 481)

the "decision will largely depend on the prospect [. . .] that [the] innovation will meet with support and sympathy from others".[233]

All disadvantages previously stated are not decisive against innovation, because each can be conceived as a price to be paid for an improvement for morality. Still, each must be hold in consideration. Thus, as Intuitionism is vague if not corrected by Utilitarianism, Utilitarianism cannot act without considering what prescribed by the Intuitional method, because it would cause more damage than happiness, which is in clear conflict with the basic axiom of Utilitarianism.

It must be noticed that, according to Sidgwick's version of Utilitarianism, even if general happiness is the paramount aim, many other secondary aims, that influence general happiness but are not general happiness, must be considered. Since this is so, some authors after Sidgwick who believe that moral actions are good or bad only if their consequences are good or bad, and believe in the importance of general happiness, but understand that many other secondary aims have to be considered, prefer to define themselves as "Consequentialists" rather than as "Utilitarian thinkers". Parfit is one of them. When Rawls writes[234] that Sidgwick's book may be said to bring a close to the period of the classical Utilitarian doctrine, is referring probably to the fact just stated.

It is worth to notice that not all Consequentialist theories are Utilitarian theories. A Consequentialist theory may aim at the best possible consequences without considering general happiness a consequence worth aiming. On the other hand, all Utilitarian theories are Consequentialist theories. For our purpose, it is not crucial to remember this distinction, since, if not specified otherwise, everything said in this work about Utilitarianism – every difficulty and every strength – will be valid also for the Consequentialist theories here considered,[235] and *vice versa*. From now on, the term "Utilitarian" and related terms will be thus replaced by the term "Consequentialist" and related terms.[236]

---

[233] (Sidgwick, 1874, p. 484)
[234] (Rawls, 1981, p. v)
[235] Not for *all* consequentialist theories, but certainly for those considered in this work. It is important to remember the distinction.
[236] The fact that the two theories are not the same theory should not cause any problem anyway. In fact, when in this work the term "Utilitarian" is used, it is intended a theory such as Sidgwick's theory, and some contemporary philosopher would probably call it a Consequentialist theory rather

We saw how Sidgwick demonstrates that Consequentialism and Intuitionism are not incompatible. In the next chapter we will examine Thomas Nagel's answer to Sidgwick's Dualism of the Practical Reason. Before that, let us restate the difficulties of Consequentialism we met so far:

- *Incalculability:* There is no precise method for comparing pleasure and pain. The *quantum* of happiness is an indefinite measurement.
- *Hedonistic Zero:* It is not clear whether the "hedonistic zero" exists and is desirable.
- *The Dualism of Practical Reason*: Rational Egoism and Consequentialism are incompatible. One of them must be proved wrong.
- *Too demanding*: Common sense believes that a moral agent should help those who are closer to her more than those who are not. Consequentialism seems not to consider this intuition. Should Consequentialism consider it? Is a method that does not consider it admissible?
- *Distributive Justice:* Is the contemporary legal system of punishment to be corrected according to Consequentialism?
- *The meaning of Utilitarianism:* Should Consequentialism be based on the *Average Principle* or on the *Total Principle*?
- *The Utility Monster*: Nozick's objection cannot be ignored. How should Consequentialism answer?

than an Utilitarian theory. In fact, Sidgwick is certainly a Consequentialist, as all Utilitarian thinkers are, but I can see no reason to consider Sidgwick less Utilitarian than Parfit.

Part 2: Recent Solutions

Chapter 5: *The possibility of Altruism*

In *The possibility of Altruism* Nagel, despite not adhering to a Consequentialist position,[237] aims to resolve the Dualism of Practical Reason. It is useful four our purpose to read Nagel's *The possibility of Altruism* for two reasons: first, as just stated, it is a first try in resolving the Dualism; second, he tries to apply Sidgwick's suggestion of using the concept of Personal Identity for challenging Egoism. Unfortunately, he will not succeed in finding a satisfying concept of Personal Identity with which solving such a Dualism. We will see, as a bonus, how Nagel demonstrates that reasons are motivating even when not prompted by a present desires.

In fact, in order to solve the Dualism, Nagel examines to what extent the motivation prompted by prudence and by altruism might have the same force by analyzing the concept of reason. Reason is not here considered as "faculty of moral cognition" as Sidgwick intended it, but as rational motive, as in the sentence "I have no reason to perform this action". Nagel tries, following Kant,[238] to find the requirements according to which "an action applies to a man on no condition about what he wants, how he feels, etc. [. . .] Requirements whose validity involves the capacity to be motivated in accordance with them".[239] Through the definition of reason he hopes to shows how, if someone has a reason to benefit herself in the future, as Prudence recommends, she has also similar reasons to benefit someone else, as Benevolence recommends, resolving the Dualism of Practical Reason. As stated, Nagel will not succeed.

---

[237] Nagel is Rawls' student, and his preferred practical solution (he proposes four solutions, finding none of them completely satisfactory) to act according to Altruism is influenced by Rawls. His favorite solution in (Nagel, 1978), that will be found further in the present work, is that, if someone wants to perform an altruistic choice, the agent should expect to live all lives influenced by that act. This is not a Consequentialist solution, since does not focus chiefly on the consequence of the action, but rather a development of Rawls' Veil of Ignorance.

[238] Nagel states to be similar to Kant in two senses. First, "it provides an account of ethical motivation which does not rely on the assumption that a motivational factor is already present among the *conditions* of any moral requirement" (Nagel, 1978, p. 13), and second because his position "assigns a central role in the operation of ethical motives to a certain feature of the agent's metaphysical conception of himself." (Nagel, 1978, p. 14). Both conditions are true also for Parfit.

[239] (Nagel, 1978, p. 12)

I will here summarize mainly the chapters in which Nagel shows how motivation comes from reason, and not from desire,[240] and sketch his reasoning about the Dualism of Practical Reason, without examining any detail but his definition of Reason.[241]

5.1. Motivating reasons

Nagel aims to show that a reason to perform an act suffices to motivate an agent even if the reason is not supported by the agent's present desires. This is a preliminary step for solving the Dualism of Practical Reason. In fact, if Nagel is right and a reason motivates regardless the support of a present desire, it would mean that, if someone succeeds in demonstrating the unacceptability of Egoism, leaving only Consequentialism and Intuitionism as the only acceptable theories, there would be no reason to act according to Egoism. Therefore, all actions suggested by the other two Methods would be motivating even if the agent feels no desire to perform the action they suggest. This would not mean that the agent would be always "accordingly motivate (if only because of the possibility of his ignorance"[242] of the action suggested by the Methods). This would simply mean that a lack of altruistic sentiment would not make less motivating a reason for acting altruistically, and therefore the agent does not need to be persuaded to perform such an act by appealing to her desires.[243] It would mean that requirements according to which "an action applies to a man on no condition about what he wants, how he feels, etc. [. . .] Requirements whose validity involves the capacity to be motivated in accordance with them"[244] can be found, and a moral action (that is an action motivated by a moral reason) can have such requirements.

The first step Nagel takes for demonstrating that reasons do not need an agent's present desire in order to motivate the agent himself is to criticize the view

[240] Those are mainly chapters V to VIII.
[241] My "sketch" of Nagel's argument is not a summary of Nagel's arguments, but more a montage of some of them.
[242] (Nagel, 1978, p. 15)
[243] Thomas Hobbes' Leviathan can be regarded as an example of a philosopher who aimed to motivate the agent to act altruistically by appealing to his egoistic desires. Nagel explains why at (Nagel, 1978, p. 9).
[244] (Nagel, 1978, p. 12)

according to which "all motivation has desire as its source".[245] The second step Nagel takes is to give a more convincing account of what a Reason is, giving his definition of Reason.

The conception that Nagel tries to confute believes that, as Nagel summarizes,

> since all motivated action must result from the operation of some motivating factor within the agent, and since belief cannot by itself produce action, it follows that a desire of the agent must always be operative if the action is to be genuinely his. Anything else, any external factor or belief adduced in explanation of the action, must on this view be connected with it through some desire which the agent has at the time, a desire which can take the action or its goal as object.[246]

This implies that, if the goal for performing an act is not related to any present desire, but only to an interest that the agent will have in the future, or to an interest that is not the agent's but someone else's interest, then the agent has no sufficient motivation to perform the act.

Nagel admits that a desire is always required in order to motivate an agent to perform an act, but he believes that it is only "a logically necessary condition. It is not necessary either as a contributing influence, or as a causal condition".[247] In order to arrive to this conclusion, Nagel begins his attack to the conception according to which all motivation has desire as its source distinguishing between motivated desires and unmotivated desires. "Unmotivated" means that it has not further motive than itself, and does not mean that is not reasonable. For example, hunger is an unmotivated desire for food. The desire to go to the supermarket once the fridge is empty is a motivated desire, and it is motivated by hunger. According to Nagel, the theory according to which all motivation has desire as its source takes for granted that all motivated desires have an unmotivated desire as motive, but this need a justification and cannot be presumed.

Let us admit that Nagel demonstrates that not all motivation comes from unmotivated desires and that, in fact, motivated desires can have as motivating condition, for example, considerations about the agent's future happiness or, as

[245] (Nagel, 1978, p. 27)
[246] Ibid.
[247] (Nagel, 1978, p. 30)

Nagel aims to show, about someone else's future happiness. If Nagel achieves that, then it has to be admitted that these considerations about the agent's future happiness, or someone's happiness, are the ones that bear the motivation to act, and that they motivate both the motivated desire and the consequent action. The motivation for performing an act would come from the reason that makes these considerations motivating, and not always from a desire. "The fact that the presence of a desire is a logically necessary condition (because it is a logical consequence) of a reason's motivating, does not entail that it is a necessary condition of the *presence* of the reason; and if it is motivated by that reason it *cannot* be among the reason's conditions."[248]

It can be demonstrated that not all motivation comes from unmotivated desires by showing that this would lead to an absurdity. It is possible to point out that, if the conception according to which all motivation has desire as its source has to be believed, it must be admitted:

(a) that the anticipation of a future desire is not *per se* a reason for a present action. In fact, unmotivated desires are presently felt (for example hunger is different from anticipation of hunger because is presently felt). An agent needs an unmotivated present desire in order to act, and anticipation of future desire is not a desire.

(b) That if someone would know that she will have a desire in the future (for example, if I know I am going to live in Poland next year, I know that next year I will desire to know the polish language), but she would have no present desire to perform an act that is necessary to satisfy that future desire (for example learning polish before going to Poland), she would have no motivation to perform that act given the truth of (a).[249]

Admitting (b) is strongly counterintuitive. A theory of motivation needs to justify the fact that a future reason can motivate even if the reason is not important for present purposes, because in our experience it is true that the anticipation of a desire provides a motivation in the present to act: if someone knew that she will live in

[248] (Nagel, 1978, p. 30)

[249] It must be noticed also that the present desire for a future object is a not *per se* a reason for a present action that is necessary to obtain the future object desired. In fact, only an anticipated future desire would provide a reason for a future object desired: the present desire cannot be guaranteed to last until the future realization of its aim if not by an anticipated future desire. But, if we believe (a), we also believe that we cannot admit future desires to be a reason for action.

Poland next year, she should feel motivated to learn polish *before* her departure, regardless any present desire to do or not do it. Therefore considerations concerning our future desires, for example concerning future desires for happiness, should be able to motivate. The theory according to which all motivation has present desires as its source have to be discharged, since it cannot give an account for acting considering the future, which is something that we know from experience we actually do.

If not only present desires can motivate, the motivation is allowed to come also from a reason that does not depend from present desires. More specifically: if I know, for example, that I will desire to be happy in the future, and I know that performing an act in the present would make me happy in the future, I would be motivated to perform an act *not* because of a present desire, but (1) because of the knowledge of a desirable end (for example the desire to be happy in the future) and (2) because of the knowledge of what should be done for achieve that desirable end (for example the performance of an act). The conjunction of (1) and (2), of desirable end and means for it, can be called "a reason" without any doubt, even if it must be more precisely defined. Nagel's more precise definition of Reason is the following:

> Every reason can be formulated as a predicate. If the predicate applies to some act, event, or circumstance (possible or actual), then there is a reason for that act, event, or circumstance to occur. Such a predicate provides reasons both primarily and derivatively: primarily, for things to which it applies, and derivatively, for things which promote that to which it applies primarily. […]
>
> Since it is people who have reasons – to act or to refrain, to promote or to prevent things – the general description of how reasons operate should show this. Let us simplify matters by (1) regarding the failure to do act B as the act of not doing B, (2) treating the prevention of A as the promotion of non-A, and (3) treating the performance of act B as a degenerate case of promoting the occurrence of act B. Then we can say that every reason is a predicate R such that for all persons *p* and events A, if R is true of A, then *p* has prima facie reasons to promote A.[250]

In the case of prudence, for all persons *p* and events A, the predicate R, said of A, is to be read as something like "A is interest of *p*", thus every person *p* has reason to

[250] (Nagel, 1978, p. 47)

promote the occurrence of A. As we saw, this kind of predicate is true both if A is interest of *p* in the present and if A is interest of *p* in the future.

In the case of altruism, for all persons *p* and events A, the predicate R, said of A, might be read as something like "A is interest of someone". In order to solve the Dualism, Nagel should demonstrate that altruistic reasons can motivate, and the most obvious way for doing it is to show that a reason R is enough to motivate every person *p* to perform an action A regardless the interest of *p*, for example by showing that *p* can understand herself to be, from some meaningful perspective, just "someone".

For showing that altruistic reasons can motivate, it must be demonstrated that reasons are invariant not only in time, but also between people. Nagel manages to do that, with the following argument. If we knew that someone is going to feel pain unless she performs an act in order to prevent it, we would think that she has a reason to perform the act that would prevent that pain, even if we do not desire to perform the act; if later we knew that *we* are who is going to feel pain unless we perform an act in order to prevent it, we would desire to perform the act, but we would not think that, simply because the disgrace happens to us (unless we perform the act), our case gives a different motivation for act than the case in which the disgrace happens to someone else (unless she performs the act). Motivation does not come only with present desires, and reasons are motivating even if someone else than us is motivated. Clearly, this is not enough for stating that we are motivated to act for benefiting someone else, as is required by Altruism. For doing it we need something more.

> In so far as a desire must be present if I am motivated to act in the interest of another, it need not be a desire of the sort which can form the *basis* for a motivation. It may, instead, be a desire which is itself motivated by reasons which the other person's interests provide. And if that is so, it cannot be among the conditions for the presence of such reasons. Desire is not the only possible source of motivation. Therefore we may look for other internal factors.[251]

Nagel thinks that those "internal factors" might be the recognition of the reality of another person and the possibility of putting ourselves in her place. In other words,

[251] (Nagel, 1978, p. 81)

we need objective reasons, and we need to conceive ourselves as someone between others, in an *impersonal* way. The weight of impersonality in Parfit will be clear in chapter 7 and 8.

Nagel underlines the importance of the conception of ourselves when we examine reasons because he believes that reasons are usually based on features of the conception that the agent has of herself. The base of what Sidgwick called *Principle of Prudence* is, as anticipated, the feature of the agent that consists in being a subject not existing only in the present, but whose existence is extended also in the future. Since this is so, it is important not to disregard the agent's interests in the future. In other words, if the predicate R is "A is *p*'s interest", R is true of A both in the present and in the future because *p* conceives herself as not only existent in the present but also in the future.

On the other hand, the *Principle of Benevolence* (that Nagel calls "Altruism") is based, according to Nagel, on the feature of the agent of being an individual among others. Parfit will try to demonstrate that, in a certain sense – an extremely relevant sense - those two features, the fact that the individual is extended in time and is one among others, are the same feature. Nagel needs to show something similar, but he cannot.[252] He guesses four ways in which, when we have to make a choice, we may conceive ourselves in a way that would help us to act altruistically. He finds none of the four satisfying,[253] but the one he prefers is the following:

> The proposal is this. To accord proper weight to the needs, desires, and interests of every individual, one must require that the choice of an interpersonal weighting principle be made under the condition that the chooser expects to lead *all* of the lives in question, not as a single super-life[254] but as a set of distinct individual lives, each of them a complete set of

[252] He is aware of the limits of his book. See (Nagel, 1978, p. 123,125)
[253] (Nagel, 1978, p. 138)
[254] As, according to Nagel, is theorized by utilitarian thinkers. According to Nagel, the theory of Utilitarianism tries to evaluate consequences from an objective point of view, as if all lives that are influenced by the agent's act are all to be considered unified in only one life, that is the one lived by the person that is evaluating the consequences. This evaluation would forget the difference of the individuals. In other words, the fact that utilitarianism tries to maximize the sum of happiness disregard the fact that, for example, the maximization may come at a great cost for an individual, and this cost may benefit only someone that is different from the individual who paid this cost. This reasoning can be found in (Nagel, 1978, p. 138), and should be probably considered valid also for many versions of Consequentialism. Parfit's version does not seem to disregard the difference between individuals. Still, his version of Consequentialism might seem to support the claim according to which it is right to burden greatly someone for someone else's benefit. To what extent it does will

> experiences and activities. If such a decision procedure could be made intelligible, it would certainly guarantee the claims of each individual a full and equal voice in the consideration of which weighting principle to adopt-a voice not merely as a possible life but as an actual one. But it is not clear how this could be done.
>
> I believe that the conditions of choice can be understood metaphysically, for we can imagine a person splitting into several persons, each of whom bears to the original, over time, the kind of relation that constitutes trans-temporal personal identity for an ordinary individual. Since identity of persons over time is not strict identity of person-stages, the loss of transitivity need not cause alarm. The relation between different stages of a single person is in some respects like the relation between different parts of the same thing. Two things can have some parts in common but not others, and two parts may bear the relation 'part of the same X' to a third part, without bearing the relation 'stages of the same person' to a third stage without bearing that relation to each other.[255]

To sum up, Nagel examines what a reason is. He finds out, after rejecting the claim according to which all motivation has desire as its source, that every reason is a predicate R such that for all persons *p* and events A, if R is true of A, then *p* has prima facie reasons to promote A. He guesses that prudence is able to provide to the agent reasons for acting due to the agent's conception of being a self existing both in the present and in the future. Since this is so, reasons prompted by prudence are invariant in time. Nagel shows that reasons can be invariant also between people, but fails to show that reasons invariant between people can motivate to act. Demonstrating the invariance of reasons between people, that is demonstrating the existence of objective reasons, is a really remarkable result, but in order to solve the Dualism of Practical Reason Nagel needs to demonstrate that the Altruism's reasons can motivate to act. He believes that objective reasons, that are *impersonal* reasons, can motivate to act if there is a metaphysical revision of the concept of the Self. Nagel fails to perform such a revision, but the suggestion of rethinking the metaphysic of the Self does not fail to be accepted by Derek Parfit, whose *Reasons and Persons* can be finally examined in this work.

---

not be matter of this work. Some hints on Parfit's response to Nagel's criticism can be found on note 340.

[255] (Nagel, 1978, p. 141)

Chapter 6: *Reasons and Persons*

The core of Derek Parfit's *Reasons and Persons* is the question with which the book begins: "What do we have most reason to do?".[256] Parfit's question coincides to some extent with our question on whether we should act according to Egoism or according to Altruism:[257] do we have more reason to benefit ourselves or others? As Sidgwick and Nagel foresaw, the answer we are looking for must be entangled with the concept of Personal Identity. Unlike Sidgwick and Nagel, Derek Parfit finds a theory of Personal Identity that allows us to solve the Dualism of Practical Reason and to state that we have more reason to follow Altruism than Egoism. Parfit's theory of Personal Identity will be analysed in chapter 7. Before that, we need to understand Parfit's strategy for answering the question on whether we should act according to Egoism or according to Altruism. He uses a military metaphor in order to explain what he is going to do. His strategy consists in finding an ally for helping Altruism to defeat Egoism. This ally is what he calls a Present-aim Theory, a fictional theory never defended by any moral philosopher, according to which what we have most reasons to do is to fulfil our present aims, and only our present aims. I will better define the Present-aim theory soon. In Parfit's own words, the strategy is the following:

> As we shall see, the Self-interest Theory [which is how Parfit calls Egoism] lies between morality and the Present-aim Theory. It therefore faces a classic danger: war on two fronts. While it might survive attack from only one direction, it may be unable to survive a double attack. I believe that this is so. Many writers argue that morality provides the best or strongest reasons for acting. In rejecting these arguments, a Self-interest Theorist makes assumptions which can be turned against him by a Present-aim Theorist. And his replies to the Present-aim Theorist, if they are valid, undermine his rejection of morality.
>
> Let us say that, in our view, a theory survives if we believe that it is rational to act upon it. A theory wins if it is the sole survivor. We shall then believe that it is irrational not to act upon this theory. If a theory does not win, having to acknowledge undefeated rivals, it must qualify its claims.[258]

[256] (Parfit, 1984, p. 3)
[257] Therefore I will examine Parfit's book only to this extent, and not in its completeness.
[258] (Parfit, 1984, p. 126-127)

Parfit will make the Present-aim Theory, supported by Morality, object Egoism and be objected by Egoism. None of the three will win neither lose. The Self-interest Theory will then need to qualify its claims. In order to qualify its claims, the Self-interest Theory must state that the most reasonable person that an agent must benefit is the agent himself or herself, conceived impartially in all parts and moments of his or her life, as prescribed by the *Principle of Prudence* (according to which, in effects, the moral agent should have impartial concern for all parts of her conscious life).

As we will see in chapter 7, according to Parfit's theory of Personal Identity the agent changes in her life enough to consider herself a different person in distant times. Not all relevant relations are equally present in all moments of a person's life, therefore there is no rational motive for benefiting ourselves in the future rather than benefiting other people. Therefore, given the validity of Parfit's theory of Personal Identity, the Self-interest Theory fails to qualify its core claim: thus, it has to be considered irrational. This is the general line of the argument, which now we have to consider more closely.

First, let us now introduce the Present-aim theory and focus on the objections that it and Morality raise towards the Self-interest Theory. The first two Arguments will not defeat any theory, but will show some interesting features of all three theories here in play.

6.1. The Present-aim Theory

The Consequentialist Theory (henceforth C) prescribes that the agent ought to cause the greatest amount of good, regardless of who is benefited by the act. According to the Self-Interest Theory, or Egoism (henceforth S), the agent ought to cause the greatest amount of good for himself, regardless of when this good will be felt.

The fictional theory Parfit names Present-aim Theory (henceforth P) states, in its most general form, that the agent ought to do what will "best achieve his present aims".[259] Aims are identified as desires intended in a broad sense, including intentions, projects and even values and ideals. This theory can assume three more specific forms, depending on what set of desires is believed worth considering. For

[259] (Parfit, 1984, p. 92)

our purpose we need to discuss here only one of these forms:[260] the *Critical Present-aim Theory* (from now on CP), according to which

> Some desires are intrinsically irrational. And a set of desires may be irrational even if the desires in this set are not irrational. For example, it is irrational to prefer X to Y, Y to Z, and Z to X. 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.[261]

The term "irrational" is used as meaning "open to rational criticism". An irrational desire might be a desire that someone wishes she did not have, or the desire to follow a pointless rule.

The form CP of P has the advantage, as we will see, that it can be translated in other theories about what we have most reasons to do, such as S or a general, unrefined theory of Altruism.

It should be clear when S and C diverge: it is when an agent can (a) benefit other people at some cost for her own interest or (b) benefiting herself at some cost for someone else. C recommends (a) at a certain extent,[262] S prescribes (b). Since bias in our own favour (that is to say the fact that we feel, instinctively, that we should benefit ourselves) are common, sometimes P may support S in prescribing (b). Particularly, a version of CP that believes that desires are irrational unless they aim at the interest of the agent coincides with S.[263] This theory is called CPS.

---

[260] I summarize the other two main forms in this note. On one form, called *Instrumental Present-aim Theory*, "what each of us has most reason to do is whatever would best fulfil his present desires" (Parfit, 1984, p. 117). By this version of the Present-aim Theory any desire provides a reason for acting. Another form is the *Deliberative Present-aim Theory*, according to which "what each of us has most reason to do is what would best achieve, not what he actually wants, but what he would want, at the time of acting, if he had undergone a process of 'ideal deliberation'—if he knew the relevant facts, was thinking clearly, and was free from distorting influences." (Parfit, 1984, p. 118). By the *Deliberative* versions of the Present-aim Theory any desire provides a reason for acting, if it survives "ideal deliberation". Further on *Deliberative* and *Instrumental Theory* can be found in (Parfit, 1984, p. 117-119)

[261] (Parfit, 1984, p. 119)

[262] That is, the agent should benefit others as far as the cost of this benefit does not outweigh the benefit itself.

[263] CPS and S coincide since in CPS any desire that does not aim at the greatest benefit to the agent has to be discarded. Parfit's emphasis on the fact that it is irrational to prefer X to Y, Y to Z, and Z to X means that a CPS-theorist cannot prefer a present lesser good to a greater later good. The present lesser good I want now can be named X, the future greater good Y: I cannot, desiring X in this

Parfit's *First Argument* against S aims at showing that that CPS, which is equivalent to S, has to be rejected. CPS has the following primary feature: according to it, the only rational desires are self-interest desires. If this claim has to be rejected, CPS has to be rejected too. If CPS and its claim have to be rejected, then S, that is equivalent to CPS, has to be rejected too.

The condition for rejecting CPS is accepting the following claim (any version of CP that accepts this claim is called (CP2)):

> *(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 in the interests of other people, when this is either morally admirable, or one's moral duty.[264]

If a version of CP accepts this claim, it conflicts with S. Here is a case in which the conflict between (CP2) and S is evident:

> *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. (In all my examples these two coincide.) 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.[265]

Parfit's considerations on this case are the following:

> On this version of CP, my act is rational. I sacrifice my life because, though I care about my own survival, I care even more about the survival of these other people. According to (CP2), this desire is no less rational than the bias in one's own favour. According to CP, given the other details of the case, it is rational for me to fulfil this desire. It is therefore rational for me to do what I know will be worse for me.[266]

---

moment, prefer now X to Y while I know that, later, I will prefer to have done Y rather than X. The difference is in formulation: while S accepts the *Principle of Prudence*, CPS aims to satisfy all the present-aims the agent will have in her life, not because of a *Principle of Prudence*, but because it would be irrational to prefer X to Y and later Y to X. According to CPS the primary aim is the present concern, whereas according to S the primary aim is the benefit of the agent. In practice, they coincide.

[264] (Parfit, 1984, p. 131)

[265] (Parfit, 1984, p. 132)

[266] Ibid.

(CP2) can be described as a general, unrefined theory of Altruism. Someone who supports S must reject (CP2). In order to show that this rejection is rational, it must be showed that (CP2) is irrational. The only way to reject (CP2) it is to state what follow:

> *The S-theorist's First Reply*: apart from the bias in the agent's favour, it is impossible that there is a desire such that the three following conditions are simultaneously met:
> 1) the desire should not be irrational
> 2) the desire should not be less rational than the bias is the agent's favour
> 3) it would be true of someone that, knowing the facts and thinking clearly, what this person most wants, all things considered, is to fulfil this desire.

In fact, if such a desire exists, there is no rational motive for rejecting (CP2).

Parfit's counterexamples for *The S-theorist's First Reply* consists in what he calls *desires for achievement*: "These are desires to succeed in doing what, in our work or more active leisure, we are trying to do. [. . .] consider artists, composers, architects, writers, or creators of any other kind. These people may strongly want their creations to be as good as possible".[267] The same is true also for scientists or philosophers that strongly want to make some discovery. It might be stated that these desires are no less rational than the bias in the agent's favour. Those people may invest time, energy and maybe wealth in their creation or discovery, well knowing that they could invest them in something that satisfies more their self-interest.[268]

[267] (Parfit, 1984, p. 133)

[268] Stating that someone might feel satisfaction in obtaining a result in her activity, and therefore the agent might be always acting according to S, misses the point. If the agent's investment is superior to the agent's satisfaction, then S would not justify that the *desire for achievement* is pursued, since S would approve an investment only if the gain is superior to the loss. This objection would provide no rational motive for rejecting (CP2). The people in the example may know that some sacrifice may improve their creation, even if this sacrifice is greater than the satisfaction that they would obtain from the creation itself. The fact that the agent would invest in her creation more than what she benefits is not irrational, if the aim is the perfection of the creation and not the satisfaction of the creator, as (CP2) allows and S does not allow. Since both the statement that the perfection of the creation is a rational aim regardless the satisfaction of the creator and its denial are defensible, this argument does not help in deciding between (CP2) and CPS, and therefore misses the point.

Unfortunately, Parfit is not able to demonstrate that these desires are no less rational than the bias in the agent's favour. But is not even possible to demonstrate its denial.

To sum up: Parfit attacks S by stating that CPS, that coincides with S, has to be rejected. Parfit proposes instead that (CP2) has to be admitted. (CP2) states that (1) benefiting the agent is not the only possible rational aim and (2) it is rational to benefit others. Parfit showed how (1) might be true, and has to demonstrate that (2) is true.

So far, CP2 and CPS (and thus also S) have to be considered equally rational. This is not a defeat or a win for none.

6.2. Self-defeating theories

An S-Theorist might state a reply different than *The S-theorist's First Reply*. An S-Theorist may reject P altogether, unless in the form that coincides with S. His *Second Reply* would be the following:

> *The S-Theorist's Second Reply*: The force of any reason extends over time. You will have reasons later to try to fulfil your future desires. Since you will have these reasons, you have these reasons now. This is why you should reject the Present-aim Theory, which tells you to try to fulfil only your present desires. What you have most reason to do is what will best fulfil, or enable you to fulfil, all of your desires throughout your life.[269]

The S-Theorist might add that P is not effective in achieving its own aims, and has to be considered self-defeating. A theory T is self-defeating if, when an agent tries to achieve the aims given by T, she achieves the T-given aims worse than how she would have achieved them by not following T. P is self-defeating because it allows that the realization of a present desire impedes the realization of a greater future desire. Therefore in certain cases S is better than P even in achieving P's own aims. P is self-defeating, unless it coincides with S, because satisfying a present desire ignoring the *Principle of Prudence* might impede to satisfy a greater future desire. Consequently, in the future, a present desire greater than the previous will not be achieved. While P is time-relative, meaning that different times have different

[269] (Parfit, 1984, p. 137)

values, S is time-neutral, since it gives all times the same importance. Time-neutrality allows S to obtain present aims more effectively than P's time-relativity.

However, *the S-Theorist's Second Reply* is weak. As Nagel guessed,[270] it might be true that the force of any reason extends not only over time, but also between different people. Furthermore, a C-theorist might claim that C is more effective than S in achieving S' own aims. Consider in fact

> *Commuters*: Each goes faster if he drives, but if all drive each goes slower than if all take buses;
> *Soldiers*: Each will be safer if he turns and runs, but if all do more will be killed than if none do;
> *Fishermen*: When the sea is overfished, it can be better for each if he tries to catch more, worse for each if all do.[271]

All other S-Theorist's objections to P can be stated also by a C-theorist against S: S is self-defeating, because obtaining a good ignoring the *The principle of Benevolence* might impede to obtain a greater good. Therefore C in certain cases is better than S even in achieving S's own aims. While S is agent-relative, meaning that different persons have different aims (agent A aims to benefit agent A, while agent B aims to benefit agent B and so on), C is agent-neutral, since it gives all persons the same aim. Agent-neutrality allows C to obtain self-interested aims more effectively than S' agent-relativity.

Neither the S-theorist's objection to P nor the C-theorist's objection to S are decisive, and therefore C has not been proved as the only rational Method so far. In fact, in 2.3 it has been described the Fundamental Paradox of Hedonism, in which it has been pointed out that the impulse towards pleasure, if too predominant, defeats its own aim. This provided no objection to any Hedonist theory, since Hedonism might be *self-effacing*. A theory T is *self-effacing* when it prescribes to the agent, in order to achieve its aim, to rely on a different theory. As stated in 2.3 Hedonism was *self-effacing*: if someone plays chess aiming only at pleasure, that is the aim of Hedonism, she will not enjoy it; if she aims to win, she will enjoy the game much more. Therefore who plays chess will achieve the aim prescribed by Hedonism by

[270] See part 5.1 of the present work
[271] (Parfit, 1984, p. 61,62)

relying on something that is *not* Hedonism. Hedonism prescribes to the agent not to rely on Hedonism in order to achieve the aims of Hedonism: this means that Hedonism is *self-effacing*. P and S might be *self-effacing* too, thus there is no motive to state that C is the best theory, yet.

Furthermore, C is not better than P and S because C defeats its own aim, too. In fact, C's aim is that consequences of an action are as good as possible. Therefore, according to C, what an agent has to do is whatever would cause the best consequences. For example, it is important for an agent to cause herself to have the desires, dispositions, beliefs, emotions and so on that are expected to allow her to cause the best consequences. Unfortunately, having the best possible set of desires, dispositions, emotions and so on might make the outcome worse.

For example, if someone has a child, the best possible set of dispositions for her includes love for the child. It is not difficult understanding how this is true: if the agent is prompted by love, will be more motivated to improve the child's life, and if things go good for the child, the love for the child will make things better also for the agent. No doubt, C would prescribe love for the child in the best set of motives of any parent.

Parfit supposes that *Clare* has one of these best sets of motives. Consider the following tragic case:

> Clare could either save her child's life, or save the lives of several strangers. Because she loves her child, she saves him, and the strangers all die.[272]

C prescribes Claire to love her child. Because of this love, Claire fails at causing the greater good in this situation.[273] C is self-defeating in this situation.

---

[272] (Parfit, 1984, p. 33)

[273] C applies to blame and praise, too. According to C, blame and praise should be allocated in order to produce the best consequences. Therefore, Claire cannot be blamed for her action: she is not guilty for causing herself to love her child and acting according to her love. Claire's act is a case of what Parfit calls *blameless wrongdoing*: the act is wrong in C's terms, but cannot be blamed, because the wrongness is caused by good dispositions. If Parfit is right, he answers successfully the objection according to which Consequentialism is too demanding: even if people would do better in sacrificing their good dispositions for a greater good, it is not mandatory to do so, and it would be paradoxical to blame them for not doing so. This means that "distinction between Excellence and Strict Duty" exists also in C, and therefore also in Utilitarianism, which is a form of C. As stated in note 231, this has been denied by Sidgwick. More on Claire and *blameless wrongdoing* can be found in (Parfit, 1984, p. 31-35)

As for S and P, the fact that C is sometimes self-defeating means not that we should reject it. But if it has to be admitted that self-defeating theories are not to be rejected, it is the case to understand *what*, then, suffices to reject a theory. Hence, it is the case to distinguish two ways in which a theory can be self-defeating. Parfit calls *indirectly individually self-defeating* theories such as the simplest version of P, according to which the agent ought to do what will best achieve his present aims. Any theory T is

> *indirectly individually self-defeating* when it is true that, if someone tries to achieve his T-given aims, these aims will be, on the whole, worse achieved.[274]

On the other hand, S and C are two *indirectly* collectively *self-defeating theories*. Any theory T is

> *indirectly collectively self-defeating* when it is true that, if several people try to achieve their T-given aims, these aims will be worse achieved.[275]

Indirectly self-defeating theories might be *self-effacing*. P, S or C might be the best theories about what we have the most reasons to do even if, for achieving their aims, they require an agent not to follow them. Consequently, *indirectly self-defeating theories* have not to be necessarily rejected. *Directly* self-defeating theories are theories that are to be rejected. Parfit calls a theory T

> *directly individually self-defeating* when it is *certain* that, if someone successfully follows T, he will thereby cause his own T-given aims to be worse achieved than they would have been if he had not successfully followed T

and

> *directly collectively self-defeating* when it is *certain* that, if we all successfully follow T, we will thereby cause the T-given aims of each to be worse achieved than they would have been if none of us had successfully followed T.[276]

[274] (Parfit, 1984, p. 5)
[275] (Parfit, 1984, p. 27)
[276] (Parfit, 1984, p. 54). Italic is mine.

Parfit shows how theories can be directly self-defeating in the First part of *Reasons and Persons*.[277] In that part he also shows how S and C might be *self-effacing*,[278] and how C cannot be directly self-defeating.[279] In those pages, Parfit achieves the surprising result of showing how the Morality of Common Sense, if not revised, is directly self-defeating.[280] Those matters are not concern of the present work.

What concern us is the following. Parfit's first argument suggested that self-interested reasons are not the paramount kind of reason. He did not achieve to demonstrate that Altruist reasons are as rational as Egoist reasons, but it was not even possible to refute both his claims.

A fictional S-theorist replied that S is better than P even in P's terms, but this objection failed, since the same can be stated by C about S. Furthermore, C is self-defeating. Once again, no theory defeated others.

Since this is so, the claims of the three theories must be qualified.

P can qualify its claim according to which an agent ought always to satisfy her present aims by stating that it is supremely rational to be concerned only about bias towards the near, that is to say caring more about the nearer future. P is agent-relative and time-relative. P is rational if bias towards the near are rational.

C can qualify its claim stating that it is supremely rational to be concerned about all consequences in time and all consequence in everyone's lives. C is agent-neutral and time-neutral. C is rational if benefiting everyone is the supreme rational aim.

S is hybrid: P is fully relative, C is fully neutral, S is agent-relative and time-neutral. S can qualify its claim stating that it is supremely rational for an agent to be concerned only about himself. S is rational if benefiting the agent is rational.

It seems that, in order to understand which claim is successfully qualified, we need to examine both time and the concept of "person". Parfit's examination of

[277] More precisely in chapter 4, (Parfit, 1984, p. 87-108).
[278] For S see (Parfit, 1984, p. 23-24), for C see (Parfit, 1984, p. 40-43).
[279] See (Parfit, 1984, p. 53-55)
[280] It can be found in (Parfit, 1984, p. 95-98). There is no reason to report this part since we have already seen in Sidgwick why a revision of Common Sense Morality is required.

"time" will not interest this work much, since S and C agree on temporal neutrality.[281]

One of the aims of this work is to show how Parfit solves what Sidgwick called *The Dualism of Practical Reason.*This dualism consists in the fact that Rational Egoism and Consequentialism are incompatible. One of them must be proved wrong.

In chapter 4 it has been reported how Sidgwick suggested to challenge the *Principle of Prudence*, according to which the moral agent should have impartial concern for all parts of her conscious life and which is the basis of Egoism, on "a ground similar to that on which Egoists refuse to admit the axiom of Rational Benevolence", according to which the good of any one individual is of no more importance than the good of any other.

Remember now how Sidgwick supposed that Hume's theories of Personal Identity may be this "similar ground": according to Hume, the Self is not a fact but a fiction, therefore no series of feelings ought to be considered of more importance than another.

Nagel tried to challenge Egoism, but he failed, since he lacked of a reliable theory of Personal Identity.

According to Parfit, also Sidgwick did not manage to resolve the Dualism of Practical Reason because he did not believe to a view similar to Hume's. Sidgwick believed to the existence of a Self that remains constant in time and that coincides with the person.[282] This view allows to challenge *the Principle of Benevolence*, since what matters for ourselves is a personal identity constant in time, therefore there is no motive to benefit others. On the opposite, the belief of a self constant in time denies any critic against *the Principle of Prudence*.

Instead, according to views like Hume's, and more precisely according to Parfit's view, Personal Identity is *not* what matters. This challenges *the Principle of Prudence*. If Personal Identity is not what matters, it is not reasonable to benefit ourselves rather than others. It cannot be claimed that, as S prescribes, the most rational aim is benefiting oneself, if oneself does not matter.

[281] The whole part 2 of (Parfit, 1984) is on Rationality and Time; further on time can be found there.
[282] Parfit's analysis of Sidgwick's "going astray" is actually more developed. The full text is in (Parfit, 1984, p. 141,142).

Let us see Parfit's examination of the concept of Personal Identity and how the common conception of the self is shaken by his analysis.

Chapter 7: Personal Identity

As there are many theories about morality, there are many theories about Personal Identity. In this part the main theories about personal identity will be briefly exposed, then it will be examined Parfit's thought experiments through which he rejects a large number of them.

The first thing to do is understanding exactly what Parfit is writing about when he examines Personal Identity. There are two kinds of identity. Two coins of the same value are *qualitatively* identical, which means that they are exactly alike. My room before and my room after I repaint are not qualitatively identical, because the room is different in the two cases. My room before and my room after I repaint it are *numerically* identical, meaning that the room is the same room even if it changed a feature. Coins of the same value are *qualitatively* but not *numerically* identical, my room before and my room after repainting it are *numerically* but not *qualitatively* identical.

According to certain views, some kinds of qualitative difference erase numerical identity. For example, it can be said that my room is numerically identical before and after its repainting, but it can be said that my room is not the same before and after its building. It might not be absurd to state that before its building my room was not even a room.

Our present examination is concerned about numerical identity of ourselves. Parfit distinguishes the following questions.

> (1) What is the nature of a person?
>
> (2) What makes a person at two different times one and the same person? What is necessarily involved in the continued existence of each person over time?
>
> The answer to (2) can take this form: '*X* today is one and the same person as *Y* at some past time if and only if ... ' Such an answer states the necessary and sufficient conditions for personal identity over time.
>
> In answering (2) we shall also partly answer (1). The necessary features of our continued existence depend upon our nature. And the simplest answer to (1) is that, to be a

person, a being must be self-conscious, aware of its identity and its continued existence over time. We can also ask

(3) What is in fact involved in the continued existence of each person over time?

Since our continued existence has features that are not necessary, the answer to (2) is only part of the answer to (3). For example, having the same heart and the same character are not necessary to our continued existence, but they are usually part of what this existence involves. Many writers use the ambiguous phrase 'the criterion of identity over time'. Some mean by this 'our way of telling whether some present object is identical with some past object'. But I shall mean *what this identity necessarily involves, or consists in*.[283]

7.1. Reductionism and Non-Reductionism

The most successful possible answers to those question are offered by what Parfit calls the Physical Criterion, the Psychological Criterion, the Dualism or *Cartesian's view* and the *Further Fact view*. As I will explain soon, in Parfit's classification, the first two theories are Reductionist, the second two are not.

The Physical Criterion is the criterion more frequently used for describing the identity of objects. It states that an object is numerically identical to another object at another time if and only if there is spatiotemporal continuity between them. For example, the painting called "the Lovers" that is currently at the MoMa museum of New York is the same that René Magritte finished painting in 1928 if a physically continuous spatiotemporal path that connects the two can be ideally traced.

This view allows to exist some things that greatly changes their features. According to the Physical Criterion a butterfly, that is an egg, then a caterpillar and finally a flying animal, is always the same being even if it changes significantly over time.

It is not clear whether this view allows gaps in the continued existence of a physical object. Parfit's example on this point involves a watch. An old man might have the same watch from his youth, but for a month it laid disassembled on a watch-repairer's shelf. On one version of the Physical Criterion, "in the spatiotemporal path traced by this watch there was not at every point a watch, so my watch does not have a history of full physical continuity. But during the month when my watch was disassembled, and did not exist, all of its parts had histories of full continuity. On

[283] (Parfit, 1984, p. 202)

another view, even when it was disassembled, my watch existed."[284]

Another controversial case concerns objects composed by parts. A famous example of this kind of cases is the ship of Theseus. Imagine Theseus' ship at the moment in which it has been constructed. After some time, a piece of the ship has to be changed. After a long time, each piece of the original ship has been changed. It is not clear if this last version of the ship is numerically identical to Theseus' ship at the moment of its construction. It must not be overlooked that the human body behaves somehow similarly to Theseus' ship: "with the exception of some brain cells, the cells in our bodies are replaced with new cells several times in our lives."[285]

The simplest version of this view, when applied to ourselves and not to objects, would state that what makes someone the same person over time is the fact that she has the same brain and body. Identity is spatiotemporal continuity of brain and body. There is a less strong version of this view, and, as usually happens, if a view is less strong, it is easier to defend. If Parfit manages to prove wrong a weak version of Psychological Criterion, also a strong version has to be rejected. In this work, for "Physical Criterion" has to be intended what follows:

> *The Physical Criterion*: (1) What is necessary is not the continued existence of the whole body, but the continued existence of *enough* of the brain to be the brain of a living person. *X* today is one and the same person as *Y* at some past time if and only if (2) enough of *Y*'s brain continued to exist, and is now *X*'s brain, and (3) this physical continuity has not taken a 'branching' form. (4) Personal identity over time just consists in the holding of facts like (2) and (3).[286]

The term "branching form" refers to particular cases, involving for examples clones and replicas, in which the subject of experiences is divided. (I will return on them with an example in 7.3.)

Let us now consider the Psychological Criterion. According to its most basic interpretation, I am not my body, but I should identify with what I think, my memory of my past experience, my attitude toward others and towards things, my intentions and other psychological traits. If I have in common with my six-years-old version the

[284] (Parfit, 1984, p. 203)
[285] (Parfit, 1984, p. 204)
[286] Ibid.

memories of my first day of school, the love for my family and the incapability of keeping my desk tidy, I am the same person as my six-years-old version. This view is rather naïve, and needs to be corrected. In fact, attitudes change, some memories are forgotten, some are added. In a reliable Psychological Criterion it is not the content of our psychological features that composes our numerical identity, but rather their continuity and connectedness. I have *not* memories and attitudes identical to my six-years-old version's memories and attitudes (even if I still love my family and cannot keep my desk tidy), but my memories and attitudes have been developed from the memories and attitudes of my six-years-old version. My psychological features are in continuity with the ones of my six-years-old version. I do not share most of my psychological traits with my six-years-old version, but I do share most of my psychological traits with my yesterday version. The distinction is defined by Parfit as follows:

> *Psychological connectedness* is the holding of particular direct psychological connections.
>
> *Psychological continuity* is the holding of overlapping chains of strong connectedness.
>
> Of these two general relations, connectedness is more important both in theory and in practice. Connectedness can hold to any degree. Between *X* today and *Y* yesterday there might be several thousand direct psychological connections, or only a single connection. If there was only a single connection, *X* and *Y* would not be, on the revised Lockean View [which is the latter, non-naïve version of the Psychological Criterion so far presented], the same person. For *X* and *Y* to be the same person, there must be over every day enough direct psychological connections. Since connectedness is a matter of degree, we cannot plausibly define precisely what counts as enough. [. . .] When there are enough direct connections, there is what I call *strong* connectedness.
>
> Could this relation be the criterion of personal identity? A relation *F* is transitive if it is true that, if *X* is *F*-related to *Y*, and *Y* is *F*-related to *Z*, *X* and *Z* must be *F*-related. Personal identity is a transitive relation. If Bertie was one and the same person as the philosopher Russell, and Russell was one and the same person as the author of *Why I Am Not a Christian*, this author and Bertie must be one and the same person. [287]

While psychological continuity is transitive, strong psychological connectedness is not. I am strongly connected to my version of an hour ago, which was strongly

[287] (Parfit, 1984, p. 206)

connected with my version of two hours ago, and so on, but I am not strongly connected with my version of ten years ago.

Since personal identity has to be a transitive relation, connectedness cannot be the criterion for establishing what personal identity necessarily involves. A supporter of the psychological criterion should turn rather to psychological continuity, and appeal to

> *The Psychological Criterion*: (1) There is psychological continuity if and only if there are overlapping chains of strong connectedness. *X* today is one and the same person as *Y* at some past time if and only if (2) *X* is psychologically continuous with *Y*, (3) this continuity has the right kind of cause, and (4) it has not taken a 'branching' form. (5) Personal identity over time just consists in the holding of facts like (2) to (4).[288]

It might be stated that it is only psychological continuity that matters. It is not. Suppose that I know that, in an hour, I will suddenly lose all my memory. Even if, after this change, the psychological continuity is preserved (since features such as my desires and affections and other psychological features are maintained), it would be very questionable to state that I did not lose a great part of my identity. In fact, I would want to avoid this change if I can, because I would not feel connected with what I will be in an hour. Imagine now that not my memory, but all my present desires and intentions will be erased in an hour: as before, my psychological continuity would be preserved, but I would try to avoid this change if I can. The same can be stated again if not my memory, not my desires but all my affections disappears, or my character is not recognisable. Despite the fact that some other features are maintained as they are, I would not feel connected with myself in an hour, and therefore my change would be perceived as a discontinuity in my identity. Therefore connectedness matters.

Depending on what is believed to be the right kind of cause for psychological continuity and connectedness, there are three versions of the psychological criterion. The *Narrow* version of the psychological criterion states that the right cause of the continuity and connectedness is the normal kind of cause. A non-normal kind of cause is this: "It is a well-established fact that people can never remember their last

[288] (Parfit, 1984, p. 207)

few experiences before they were knocked unconscious."[289] If I fall unconscious and later someone tells me some likeable detail of what happened just before I fell unconscious, I might seem to remember this experience. Even if I actually had that experience, I have not it in my memory with the right kind of cause, because this (apparent) memory is not caused by my experience but by someone else's account.

The same applies for other kinds of continuity, such as continuity of character. If character is changed for normal causes, such as changes strongly willed by the changing person, or natural consequences of growing older, or natural responses to certain experiences, the *Narrow psychological criterion* will describe the continuity as not interrupted. On the other hand, "there would not be continuity of character if radical and unwanted changes were produced by abnormal interference, such as direct tampering with the brain."[290]

The *Narrow Psychological Criterion* coincides in most cases with the Physical Criterion: the normal cause of psychological continuity involves the continued existence of enough of the brain. Though brain's existence is not a sufficient condition for the normal cause of psychological continuity, it is a necessary condition.

The other two versions of *The Psychological Criterion* are the *Wide* and the *Widest* version. On the *Wide* version of the psychological criterion the right cause of the continuity is any reliable kind of cause. On the *Widest* version, any cause of psychological continuity is the right cause.

Parfit rejects immediately the *narrow* version.

> A partial analogy may suggest why. Some people go blind because of damage to their eyes. Scientists are now developing artificial eyes. These involve a glass or plastic lens, and a micro-computer which sends through the optic nerve electrical patterns like those that are sent through this nerve by a natural eye. When such artificial eyes are more advanced, they might give to someone who has gone blind visual experiences just like those that he used to have. What he seems to see would correspond to what is in fact before him. And his visual experiences would be causally dependent, in this new but reliable way, on the light-waves coming from the objects that are before him.

[289] Ibid.
[290] Ibid.

> Would this person be *seeing* these objects? If we insist that seeing must involve the normal cause, we would answer No. But even if this person cannot see, what he has is just as good as seeing, both as a way of knowing what is within sight, and as a source of visual pleasure. If we accept the Psychological Criterion, we could make a similar claim. If psychological continuity does not have its normal cause, it may not provide personal identity. But we can claim that, even if this is so, what it provides is *as good as* personal identity.[291]

Parfit will reject also the *Wide* version of the *Psychological Criterion*, as I will show in 7.4; before that, it is worth underlying that both the *Physical* and *Psychological Criteria* are Reductionist views about Personal identity. The Reductionists claim:

> (1) that the fact of a person's identity over time just consists in the holding of certain more particular facts.

They may also claim

> (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. These facts can be described in an impersonal way.[292]

It must be noticed that according to a Reductionist view what is necessarily involved in a person's continued existence is less than what is in fact involved. All Reductionist views agree in stating that

> (3) A person's existence just consists in the existence of a brain and body, and the occurrence of a series of interrelated physical and mental events.

Some Reductionists claim

> (4) A person just is a particular brain and body, and such a series of interrelated events. [293]

[291] (Parfit, 1984, p. 208-209)
[292] (Parfit, 1984, p. 210)
[293] This is the Physical view.

Other Reductionists claim

> (5) A person is an entity that is distinct from a brain and body, and such a series of events.[294]

All Reductionists agree on (3), but believe that what is involved in the existence of a person is either (4) or (5); both are *less* than what stated in (3). (5) deserves a closer look, which we can provide by appealing to the following Hume's analogy:

> I cannot compare the soul more properly to anything than to a republic, or commonwealth, in which the several members are united by the reciprocal ties of government and subordination, and give rise to other persons, who propagate the same republic in the incessant changes of its parts. And as the same individual republic may not only change its members, but also its laws and constitutions; in like manner the same person may vary his character and dispositions, as well as his impressions and ideas, without losing his identity.[295]

A Reductionist about nations (that is the most common view about nations) would accept claims analogous to (3), (4) and (5). Those claims are

> (6) A nation's existence just involves the existence of its citizens, living together in certain ways, on its territory.

Some claim

> (7) A nation just is these citizens and this territory.

Others claim

> (8) A nation is an entity that is distinct from its citizens and its territory.[296]

All Reductionists about nations agree on (6), but believe that what is involved in the existence of a person is either (7) or (8); both are *less* than what stated in (6).

Claim (2) means that we could give a *complete* description of a persons

[294] (Parfit, 1984, p. 211). This is the Psychological view.
[295] (Hume, 1740, p. 273)
[296] (Parfit, 1984, p. 211-212)

without claiming that a person exists. For example, a reductionist can state that there exist a particular brain and body, and a particular series of interrelated physical and mental events. If a Reductionist accepts (4), this is *equivalent* to state that a person exists, tough it has not been claimed that this person exists. If a Reductionist accepts (5) the fact that there exist a particular brain and body, and a particular series of interrelated physical and mental events is not equivalent to stating that a person exists, but it *implies* that a person exists. Reductionists that accept (5) can state that, if a description "of reality either states or implies, or enables us to know about, the existence of everything that exists, our description is complete."[297] This statement is plausible.

In order to understand how a description might be complete even without claiming that something exists, and how, if a description enables us to know about the existence of everything that exists, the description is complete, it will help to report Parfit's example.

> Consider, for example, clubs. Suppose that a certain club exists for several years, holding regular meetings. The meetings then cease. Some years later, some of the members of this club form a club with the same name, and the same rules. We ask: 'Have these people reconvened the *very* same club? Or have they merely started up *another* club, which is exactly similar?'[298]

Suppose that the original club might have had no rule explaining how, after such a period of non-existence, it could be reconvened, or a rule preventing it. Our question would be an empty question: we know *all* the facts, and we miss none. There is nothing more to know. The description is complete. We can *decide* to give an answer to this question, but it is not a decision between different views about what happened and what exists. Different views about what happened are simply different possible interpretations of a fact, but what matters is the *fact*, not the interpretations. If we know everything about a fact, and miss nothing about it, we need not to ask further questions. Any further question would be empty.

[297] (Parfit, 1984, p. 213)
[298] Ibid.

Similarly, Reductionists might answer to some question such as "will the resulting person of this person be me? Am I about to die?" stating that this question has no answer. For the moment, this will certainly seem hard to believe. The belief that our identity must always be determinate holds in a great number of cases. We will see cases in which the most plausible belief will be that a question as "am I about to die?" is empty. A question is empty when, despite knowing all the facts concerning that question, and missing none, we still cannot answer to that question. As will be shown later, sometimes answering to the question "am I about to die?" would be merely give an interpretation to a fact and, as stated above, what matters are the facts, not the interpretations.

Other kinds of theories about Personal Identity are Non-Reductionist. A theory is Non-Reductionist when it rejects those claims previously stated:

> (1) that the fact of a person's identity over time just consists in the holding of certain more particular facts.
> (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. These facts can be described in an impersonal way.[299]

A Non-Reductionist view believes that we are separately existing entities. Personal Identity does not involve physical or psychical continuity: according to this view identity consists is something distinct from brain and body and experiences. "On the best-known version of this view, a person is a purely mental entity: a Cartesian Pure Ego, or spiritual substance. But we might believe that a person is a separately existing physical entity, of a kind that is not yet recognised in the theories of contemporary physics."[300] This view is called Dualism, or *Cartesian view*.

Another Non-Reductionist view denies that a person is a separately existing entity, but even so a person cannot be identified just through her psychological and physical features. According to this view, personal identity is a further fact. This view is hence called *Further Fact View*.

[299] (Parfit, 1984, p. 210)
[300] Ibid.

So far, views about the nature of Personal Identity have been described. Parfit adds views about the *importance* of Personal Identity.

> Consider an ordinary case where, even on any version of the Reductionist View, there are two possible outcomes. In one of the outcomes, I am about to die. In the other outcome I shall live for many years. If these years would be worth living, the second outcome would be better for me. And the difference between these outcomes would be judged to be important on most theories about rationality, and most moral theories. It would have rational and moral significance whether I am about to die, or shall live for many years. What is judged to be important here is whether, during these years, there will be someone living who will *be me*. This is a question about personal identity. On one view, in this kind of case, this is always what is important. I call this the view that *personal identity is what matters*. This is the natural view.
>
> The rival view is that *personal identity is not what matters*. I claim
>
> > *What matters is Relation R*: psychological connectedness and/or continuity, with the right kind of cause.
>
> Since it is more controversial, I add, as a separate claim
>
> > In an account of what matters, the right kind of cause could be any cause.[301]

Parfit believes that (a) our Identity over time is determined by Relation R, if it does not take a "branching" form (It will be clear what does "a branching form" means in 7.3). Parfit supports the *Widest Psychological Criterion*. He also believes that (b) our identity is not always determinate, and that (c) Personal Identity is not what matters, but Relation R is all that matters. He will also argue that, (d) from our birth to our death, what we conceive to be ourselves are actually a series of successive selves. His arguments will be here exposed in this order: in 7.2 I am going to expose an argument according to which the *Further Fact View* is rejected and (b) is necessarily implied by Reductionism; in 7.3 I will summarize two arguments for accepting (c) and rejecting non-Reductionist views (and therefore accepting (b), implied by Reductionism as showed in 7.3); in 7.4 I will introduce an argument for rejecting the Physical Criterion (and therefore accept the only remaining claim about Personal Identity, which is (a)). The exposition will be concluded with the argument for (d). If (d) is true, S cannot qualify its claim, and has to be considered irrational.

[301] (Parfit, 1984, p. 215)

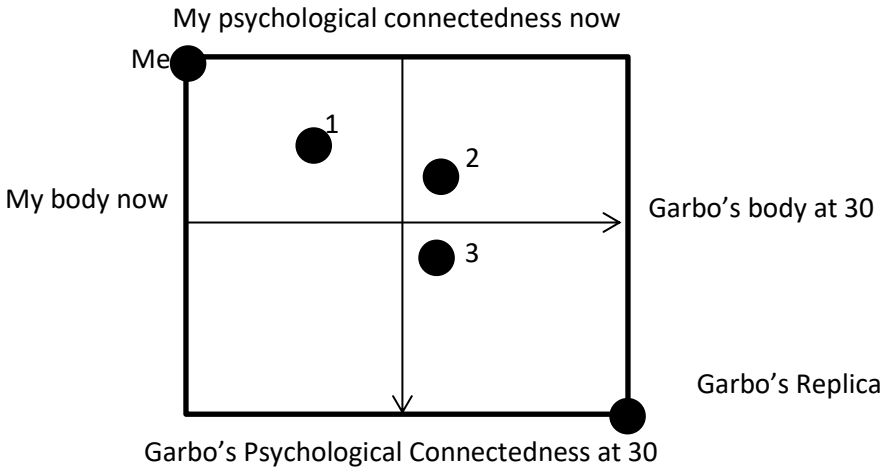
7.2. The Psychophysical Spectrum

Following Parfit's train of thought, it is possible to challenge now the view according to which our identity is always determinate. Consider what I call

> *The Psychophysical Spectrum*.[302] I am the prisoner of some callous surgeon, who intends to disrupt my psychological and physical continuity by tampering with my brain and body. Suppose that he knows how to perform all the alterations he wants without interrupting my biological functions, and therefore without killing. I shall be conscious while he operates, and in pain: I will be in pain regardless how much he tampers with my brain and body. For example, he might decide to open my head without touching my brain, and I will be in pain, since my head is open. I therefore dread what is coming. The surgeon tells me that, while I am in pain, he will apply on my brain, by flipping some switches of a first set of switches, a certain number of neurodes. The greater the number of neurodes applied, the more my psychological connectedness, which consists in features such as my memory and character, will be erased and substituted with the psychological connectedness that Greta Garbo had in her 30th birthday, that the surgeon recorded in a previous experiment. Any particular switch would cause only a very small change. If few switches are flipped I will maintain most of my features. If all switches will be flipped, all my psychological features will be replaced by Greta Garbo's. The fact that my psychological connectedness might overlap to some extent with Greta Garbo's gives me less reason to dread what is coming? Can I assume that, if the surgeon flips all the switches, my pain will suddenly cease? The pain might so occupy my mind that I would even fail to notice the loss of all these features.
>
> The surgeon next tells me that, while I am still in pain, he will later flip some switches of another set of switches, that will replace the cell of my body with cells that have the features of Greta Garbo's cell at the age of 30. In fact, the surgeon recorded also the state of her cells in the previous experiment I already mentioned. Again, any switch produces only small changes and the more switches will be flipped, the more my cells will be replicas of Greta Garbo's cells. Can I assume that this will cause my pain to cease?

The following square scheme covers all the spectrum of possible cases.

[302] This argument is never exposed as such in *Reasons and Persons*, but it is a summary of all arguments in (Parfit, 1984, p. 229-243). Parfit's arguments are a development of a Bernard Williams' argument present in 'The Self and the Future', *Philosophical Review* 79, No. 2, Apr. 1970 . Parfit calls *The Combined Spectrum* the argument which is most similar to the one I write there, because it combines a *physical* and a *psychical spectrum.* I call it *The Psychophysical Spectrum* so that it is clear what features of the individual are involved. What I call *The Psychophysical Spectrum* is therefore a montage and an adaptation of Parfit's argument in (Parfit, 1984, p. 229-243).



The person I am *before* the experiment is located at the square's top-right corner. The person that will feel pain *during* the surgeon's experiment is a spot somewhere in the square. The more switches of the first set of switches will be flipped by the surgeon, the closer the spot will be to the square's bottom. The more switches of the second set of switches will be flipped by the surgeon, the closer the spot will be to the square's left.

If the surgeon flips no switch, I will be the person that feels pain. If the surgeon flips all the switches, a Replica of Greta Garbo will be the person who feels pain. Those who believe a Reductionist theory about Personal Identity would state that, if all switches are flipped, I will be dead, since both my psychical and physical continuity have been broken. If the surgeon flips one switch, the person feeling pain can be described as me with, for example, slightly better acting abilities. Since this difference in my personality is trivial, many people would state that I am not dead and I am the person feeling pain during the operation; certainly, I should dread what is coming. Suppose then than the surgeon flips a second switch, causing a second slight change in my physical or psychical features. Am I alive, am I in pain? Probably yes, I am, since I change, once again, trivially. And with a third switch flipped, with the consequent third slight change? What about with enough switch flipped to reach spots 1, 2 or 3? What about all switches?

It would be hard to believe both that I will survive or die if the surgeon will flip one switch less or one switch more. It seems that whether I will be alive or dead,

whether I will or I will not be the person feeling pain, cannot depend on very small changes on my memories, or on my personality, or on my body. But, if no small change would cause me cease to exist, I would be the person feeling pain in all possible cases of the spectrum, included the case in which all my physical and psychical features coincides exactly with Greta Garbo's features at the age of 30. This is absurd.

It might be stated that we can resolve this matter with a stipulation. If we rely in the Psychological Criterion, we might agree, for example, that I am alive if at least 50% of my psychological features are maintained. If we rely in the Physical Criterion, we might agree that I am alive with at least 50% of my physical features. But we cannot decide that I am alive or that I am dead with a *stipulation*. There must be a *relevant* argument for believing that I am alive or that I am dead. The criteria for Personal Identity cannot be stipulated. If we stipulate the line of division between my life or my death, we cannot believe that this line has any intrinsic value. Drawing a line between two neighbouring cases, though the difference between them is trivial in itself, is an arbitrary operation.

Appealing to what thinks who will feel the pain cannot help. Somewhere in the square there is certainly a point after which a person believes or does not believe to be me. But we have no reason to trust this belief, since this person's mind is tampered by the surgeon.

A Reductionist might claim that, in central cases of the squares, there is no answer to the question if I am about to die, or if I am the person that is feeling pain. But I know exactly what is happening, and I can make a complete description of this case. Questions about whether I am about to die might be empty questions. A question is empty when we know all the facts, and miss none, even if we could not answer to that question. Since in each case it is possible to know how much I will be connected to the person that suffers by the surgeon's hand, I know everything. It is possible that, if the suffering person will be in the central spots of the square, I can describe myself both as surviving in pain the surgery or as dead in the surgery. Neither of the two descriptions would be true or false: "those would be merely two

possible description of the very same course of events".[303] Even ignoring if I will be dead or alive, I would know everything there is to know.

We have three possibilities.

(1) Accepting the Reductionist claim just stated.

(2) Believe that there is a sharp borderline that separates cases in which I am alive and I am dead, *"even though there could never be any evidence where these cases are"*.[304]

(3) Accepting the *Cartesian view* of Personal Identity.

It is hard to believe (a) that the difference between life and death could consists in very small differences of some of my features. It is also very hard to believe that (b) there must be a borderline that divides cases in which I will live and I will die, but we could never know where it is. It might be possible to accept one between (a) or (b). (2) requires to believe in *both* (a) and (b). Believing in both (a) and (b) is extremely implausible, or at least more implausible than believing in (1).

Notice that *the Psychophysical Spectrum* provides a very strong objection to the *Further Fact View*. According to this view, a person cannot be reduced to her physical or psychical features, but consists also in a further fact. What is this further fact? Is there a further fact that can justify that, after the surgeon has flipped all the switches, *I* am a person *exactly* similar, in all features, to Greta Garbo at the age of 30? Or, if there is not, how could I decide whether or not I will be alive or dead? "If we are not separately existing entities, in what could this further fact consist? What could make this fact [in the central cases of the spectrum] either hold or fail to hold?"[305]

The possibilities that can be accepted are (1) or (3). It must be noticed that Reductionists can accept only (1). This means that Reductionists must admit that identity sometimes can not be determinate, and sometimes questions such as "am I about to die?" are empty questions. The Psychophysical Spectrum shows how all Reductionists that refuses to believe that identity sometimes can be indeterminate are not coherent.

[303] (Parfit, 1984, p. 233)
[304] (Parfit, 1984, p. 239)
[305] (Parfit, 1984, p. 240)

We might instead accept (3), the *Cartesian View*, according to which we are entity that exist separately from our psychophysical features. Unfortunately, "there is no [scientific] evidence in favour of this view, and much evidence against it."[306] The most striking evidence against the *Cartesian View* and therefore in favour of Reductionism does not come from a thought experiment; it comes from an actual experiment concerning humans suffering of a particular kind of epilepsy. The most important articles that report and comment those experiment is (Nagel, 1971), for philosophers, and (Sperry, 1966), for those interested in anatomy and neurosciences.

7.3. Brain bisection

Nagel's and Sperry's articles are very detailed, and both can be read without any difficulty also by non-specialists. For a full report of the facts I recommend reading them. In what follows I will report only what is needed for the purposes of this work.
The top of the human brain has two hemispheres, with different functions and linked with the opposed side of the body (the right hemisphere is linked with the left side of the body and *vice versa*). The two hemispheres communicate with each other through a band of nerve fibres called corpus callosum and trough smaller pathways. In a normal person, the corpus callosum and those pathways help to coordinate and integrate the functions of the two hemispheres. In order to cure a certain kind of epilepsy, in the Britain of 1950 the corpus callosum and the smaller pathways between the two hemispheres were surgically severed. Those who were affected by epilepsy and were treated in this manner were successfully cured. Those patient seemed to behave normally in their everyday life. But Sperry and Cazzaniga discovered that, if the visual field of those person was separated, the right side of the brain (which govern the left side of the body) and the left side of the brain (which govern the right side of the body) behaved as "two cognitive entities",[307] or "two free wills packed together inside the same cranial vault".[308]

> Here is a simplified version of the kind of evidence that such tests provide. One of these people is shown a wide screen, whose left half is red and right half is blue. On each half in a darker shade are the words, 'How many colours can you see?' With both hands the person

[306] (Parfit, 1984, p. 243)
[307] (Sperry, 1966, p. 302)
[308] (Sperry, 1966, p. 304)

> writes, 'Only one'. The words are now changed to read, 'Which is the only colour that you can see?' With one of his hands the person writes 'Red', with the other he writes 'Blue'. If this is how this person responds, there seems no reason to doubt that he is having visual sensations—that he does, as he claims, see both red and blue. But in seeing red he is not aware of seeing blue, and vice versa. This is why the surgeon writes of 'two separate spheres of consciousness'.[309] In each of his centres of consciousness the person can see only a single colour. In one centre, he sees red, in the other, blue.
>
> The many actual tests, though differing in details from the imagined test that I have just described, show the same two essential features. In seeing what is in the left half of his visual field, such a person is quite unaware of what he is now seeing in the right half of his visual field, and vice versa. And in the centre of consciousness in which he sees the left half of his visual field, and is aware of what he is doing with his left hand, this person is quite unaware of what he is doing with his right hand, and vice versa.[310]

Such conflict may take forms that are more sinister. The Nobel prize Robert W. Sperry reports that

> [A] patient and his wife used to refer to the "sinister left hand" that sometimes tried to push the wife away aggressively at the same time that the hemisphere of the right hand was trying to get her to come and help him with something.[311]

Those people *never* experience that their consciousness is divided. Nevertheless, when their visual field is separated, they behave differently. While a mind is divided, there are at least two subjects of experience, that have different abilities, and different thoughts. Their consciousness splits when the visual field splits. It is extremely likely that every human, if her brain and visual field is split, would behave exactly as "two cognitive entities." Parfit explains how we can understand this fact with this analogy: "We can come to believe that a person's mental history need not be like a canal, with only one channel, but could be like a river, occasionally having separate streams."[312]

How would an holder of the *Cartesian View* explain these facts? How does she describe, in her theory's terms, the fact that there are *two* subjects of experience

[309] (Sperry, 1966, p. 299)
[310] (Parfit, 1984, p. 245)
[311] (Sperry, 1966, p. 304)
[312] (Parfit, 1984, p. 247)

and the fact that the person with the divided brain is *neither of them*? An holder of the *Cartesian View* cannot claim that the *Cartesian Ego* is the person that has the divided brain, since (1) she is not the subject of the experiences and (2) there would be subjects of experiences that are not *Cartesian Egos.* What would they be, then?

An holder of the *Cartesian View* cannot even claim that the *Cartesian Ego* is one of the two persons, because, once again, there would be subjects of experiences that are not *Cartesian Egos*.

Those who believes in the *Cartesian View* cannot explain the phenomena involved in the division of the brain. Since those phenomena are real, a view that cannot explain a real phenomenon cannot be considered adequate.

The only explanation for the case of the split brains, unacceptable by non-Reductionists, seems be to be the following: people are usually conscious of more things contemporaneously. I am conscious of the sound of my fingers on the keyboard, of the light that comes out from the screen, of the solidity of my chair and so on. When our brain is split, we are having different sets of experiences for each hemispheres. When our brain is not split, we have only a set of experience.

> 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. It is worth restating other parts of the Reductionist View. I claim:
>
> > Because we ascribe thoughts to thinkers, it is true that thinkers exist. But thinkers are not separately existing entities. The existence of a thinker just involves the existence of his brain and body, the doing of his deeds, the thinking of his thoughts, and the occurrence of certain other physical and mental events. We could therefore redescribe any person's life in impersonal terms. 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. Persons would be mentioned here only in the descriptions of the content of many thoughts, desires, memories, and so on.[313] Persons need not be claimed to be the thinkers of any of these thoughts.
>
> These claims are supported by the case where I divide my mind. [. . .] There are only two alternatives. We might ascribe the experiences in each stream to a subject of experiences

[313] The person would in fact be mentioned in self-referring thoughts, memories and so on.

> which is not me, and, therefore, not a person. Or, if we doubt the existence of such entities, we can accept the Reductionist explanation. At least in this case, this may now seem the best explanation.[314]

There is one more feature of our brains that is worth noticing: "an intact brain contains two cerebral hemispheres each of which possesses perceptual, memory, and control systems adequate to run the body without the assistance of the other".[315] Furthermore, "there are many people who have survived, when a stroke or injury puts out of action one of their hemispheres".[316] This allows Parfit to write another thought experiment, that will be useful in order to demonstrate that identity is not what matters.

Since all the views defined by Parfit as Non-reductionist has been rejected, it can be stated that a person would survive if her brain, together with her consciousness, was successfully transplanted into a similar body – let us say, in a twin's body, in order to be sure that the transplanted brain will not lose part of his psychological connectedness due to significant physical changes, such as a change of sex. Furthermore, for the motive stated in the previous paragraph, a person could survive even if only *one* hemisphere is transplanted, the other hemisphere having been destroyed.

But what if the other hemisphere has *not* been destroyed? Assume that I am one of three identical triplets. Consider

> *My Division*. My body is fatally injured, as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes that he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me. And he has a body that is very like mine.[317]

---

[314] (Parfit, 1984, p. 251)
[315] (Nagel, 1971, p. 410)
[316] (Parfit, 1984, p. 254)
[317] (Parfit, 1984, p. 254-255) Notice that this case is likely to remain impossible, for several reasons. For example, in order to perform an operation of this kind, it is necessary to severe not only the two hemispheres, but also the lower brain, which is unlikely to be divided without killing the owner of the brain. Further detail on the impossibility of such a surgery can be found at (Parfit, 1984, p. 255). The technical impossibility of this surgery is nevertheless trivial. "The one feature of the case that might be held to be deeply impossible—the division of a person's consciousness into two separate streams—is the feature that has actually happened" (Parfit, 1984, p. 255). It *is* possible for someone to survive with only half brain, and there *are* two spheres of consciousness. It is true that it is anatomically

I am continuous, both physically[318] and psychologically, with both. This form of continuity is an example of what is called by Parfit a "branching" form. Both of the resulting people are fully psychologically continuous with me. What happens to me?

The possibilities are: (1) I am dead; (2) I survived as one of my twins; (3) I survived as the other; (4) I am both my twins.

It would be paradoxical to believe (1). Remember that, according to the thought experiment we are asked to consider, we hypothesize that a human being can survive if half of her brain is transplanted,[319] the other half having been destroyed. It would be paradoxical to state that, since the other half has not been destroyed, the human being is dead. Two successful surgeries cannot result in the death of the patient, while a success and a failure results in her survival.

About (2) and (3), we have no ground for supporting neither one neither the other. I am continuous with *both*. The two halves of the brain are mine at the same extent. It is not plausible to choose one of the two, because it would mean to discard the other, which has the same exact features of the former, and therefore the same exact rights to be chosen as the half in which I survive.

Since both halves have the same rights to be chosen as the surviving half, we might think that (4) is the correct answer. But, if I was only a person *before* the surgery, how can I be two people *after* the surgery? Any possible interpretation involves an implausible distortion of the concept of person.

On one interpretation of (4) my twins are the same person: me. Imagine that

> After I have had this operation, the two 'products' each have all of the features of a person. They could live at opposite ends of the Earth. Suppose that they have poor memories, and that their appearance changes in different ways. After many years, they might meet again, and fail even to recognise each other. We might have to claim of such a pair, innocently

---

impossible to divide the spheres in order to make them rule two different bodies, but it is nevertheless worth to consider the case in which it might happen. In the same way, it is impossible to know what we would see if we could travel beside some beam of light at the speed of light, but it is still worth asking it, or at least it was worth asking for Einstein, as Parfit reports in (Parfit, 1984, p. 219)

[318] Remember that for physical continuity it is required to have enough of the brain, and not the whole body. Since I survived, enough of the brain is continuous in both the "products" of the transplants.

[319] I remark that this is possible only with a single hemisphere, not with half brain; see note 317.

> playing tennis: 'What you see out there is a single person, playing tennis with himself. In each half of his mind he mistakenly believes that he is playing tennis with someone else.' [320]

Are the two tennis-players a single person? Either this question is empty, or the answer is No. "It cannot be true that what I believe to be a stranger, standing there behind the net, is in fact another part of myself."[321]

On another interpretation of (4) my twins are different persons, but yet they are me, "in the way in which the Pope's three crowns together form one crown".[322] But "Suppose the resulting people fight a duel. Are there three people fighting, one on each side, and one on both? And suppose one of the bullets kills. Are there two acts, one murder and one suicide? How many people are left alive? One or two?"[323] This view is too implausible, or at least less plausible than believing that the question "who is me after the division?" is empty.

Once again, we are forced to the conclusion that our identity is not always determinate, and that the question "am I dead or alive?" is an empty question. We can describe the outcome of the operation in different ways, all of which is neither true nor false. The possibility of an empty question in Personal Identity has been discussed in 7.2. The further claim that can be done thanks to *My Division* is that *it does not matter* which description is better, and if I am one or two or three persons after the division. What matters is something else. What matters matters for us – or for *me*, since *I* am the one that undergoes this operation. Should I regard the prospect of the division as like death or as like survival? This is what matters.

My relationship with the two "products" of the transplants is of psychological and physical continuity. This relationship is the same relationship that I have with myself during time in ordinary cases. My relation with the "products" contains everything needed for me to survive as a person, and nothing is missing. There is no ground for believing that the nature of my relationship with the two "products" of the division is, for me, as bad as death. Since the duplication is not as bad as death, it is as good as ordinary survival. The division is not the same as ordinary survival, but is not

[320] (Parfit, 1984, p. 256)
[321] (Parfit, 1984, p. 257)
[322] Ibid.
[323] Ibid. There are other possible interpretations of (4), which leads to even more implausible distortion of the concept of person. I will not comment them here. They can be read at (Parfit, 1984, p. 257-258)

nearly the same as death: stating that division is death is like "confuse two with zero".[324] The only difference between ordinary survival and the division is that the latter does not fit the logic of identity, but *identity is not what matters*. It matters what becomes of ourselves.

Parfit states therefore that

> *Relation R* is what matters. R is psychological connectedness and/or psychological continuity, with the right kind of cause.
> In an account of what matters, the right kind of cause could be any cause.[325]

Parfit writes*:*

> A future person will be me if he will be R-related to me as I am now, and no different person will be R-related to me. If there is no such different person, the fact that this future person will be me just consists in the fact that relation R holds between us. There is nothing more to personal identity than the holding of relation R.[326]

In a great majority of actual cases[327] only one person is R-related to a future or past self, therefore Personal Identity coincides with the relation R. In cases like *My Division* they do not coincide. In those cases, the only things that are connected or continue are Relation R, and part of my brain. Suppose that, as will be demonstrated in 7.4, the brain is not what matters. Since in *My Division* it cannot be stated that a person dies, Relation R is enough for me to state that I do not die. This claim supports the two wide *Psychological Criteria.*

If we have to decide what matters to us, we have to decide between Personal Identity and Relation R. If Reductionism is right, if Personal Identity is not a deep further fact and if we succeed in demonstrate that the Physical Criterion has to be rejected, we must accept that what matters is Relation R. In fact, in the case of *My Division* asking if Personal Identity is preserved is an empty question. But I, that is to say the person that undergoes *My Division*, do not stop to care about what happens after the transplant: I still care about the future, and think that something matters.

[324] (Parfit, 1984, p. 262)
[325] Ibid.
[326] Ibid.
[327] Some organism, like the amoebae, make exception.

What is the state of my Personal Identity in this case is not what matters: what matters is the state of my Relation R. In the case of *My Division*, once someone understands *why* neither resulting person would be her, this person also sees why it does not matters: all what matters is how the Relation R will be.

It might be stated that actually the loss of the unity of consciousness, even if it is not as bad as death, it is still *less* good than ordinary survival. In practice, the fact that what once was a person is now two persons may cause problems: for example, the wife or the husband of the divided person may have some trouble in understanding if his or her love must be given to a copy, to the other, to both, or to none. But those kind of practical problems does not make the case of the division less good than ordinary survival, since they are merely contingent problems. There might be contingent advantages, since each "product" of the operation has the possibility to collaborate with someone that has an incredible affinity with her. Since advantages and disadvantages are contingents, there is no argument for stating that loss of the unity of consciousness is *less* good than ordinary survival.

7.4. Successive Selves and the Dualism of Practical Reason

If Personal Identity is not what matters, but what matters is Relation R, it is not true that an agent's supreme rational aim is to benefiting herself. In fact, if who oneself is does not matter, it cannot be stated that oneself should be privileged: it cannot be stated that benefiting something that does not matter is supremely rational.

The claim that the supreme rational aim is to benefiting oneself is the basic claim of Egoism. Since Egoism's central claim ought to be rejected (which implies that this claim cannot be qualified), and since Egoism has not been proved superior to P and C in chapter 6, Egoism should be rejected.

But an Egoist might be not convinced yet that Personal Identity does not matter. An Egoist believes that it is supremely rational to be concerned only about her own self-interest. What relation would an Egoist think that would fundamentally matter?

> If the rest of [a] person's life will be well worth living, in what way should [an Egoist] want to be related to this person? [. . .] This relation will also be what, for all of us, should fundamentally matter, in our concern for our own future. But since we may be concerned

> about the fate of the resulting person, whatever his relation is to us, it is clearest to ask what, for an Egoist, should matter. Here are the simplest answers:
> (1) Physical continuity,
> (2) Relation R with its normal cause,
> (3) R with any reliable cause,
> (4) R with any cause.
> R is psychological connectedness and/or continuity, with the right kind of cause. If we decide that R is what matters, we must then consider the relative importance of connectedness and continuity. It might be suggested that what matters is both R and physical continuity. But this is the same as answer (2), since physical continuity is part of R's normal cause.[328]

The normal cause of the Relation R is the continuous permanence in time of a brain and a body, that belongs to the same person.

A reliable cause of Relation R is, simply, any cause on which we can rely. For example, the normal cause for sight is the naked eye; a reliable cause for sight is an eye helped by the lens of glasses.

Stating that any cause is good for the holding of Relation R means that any cause is a reliable cause.

To defend (1) we need to believe that, if someone shall be physically continuous with some resulting person, this is what matters, even if she shall not be R-related to this person.

Suppose that I know that, somehow, a surgeon destroys my psychological continuity, leaving my body intact. There is no way to restore my psychological continuity. If I was an Egoist, should I be concerned for this future individual? Remember that, according to the Physical Criterion, not all body must be preserved in order to guarantee Personal Identity. If all the body should be preserved, people that has been cured through a transplant or a prosthesis of any kind would have lost their identity for this reason. According to the Physical Criterion, in order to guarantee Personal Identity, it suffices that enough of the brain is preserved. Parfit writes:

> Why should the brain be singled out in this way? The answer must be: 'Because the brain is the carrier of psychological continuity, or Relation R'. If this is why the brain is singled out,

[328] (Parfit, 1984, p. 283)

> the continuity of the brain would not matter when it was not the carrier of Relation R. The continuity of the brain would here be no more important than the continuity of any other part of the body. [. . .] If R will not hold, the continuity of the brain should have no significance for the person whose brain it now is.[329]

This is a strong argument in favour of *The Psychological Criterion*, and against *The Physical Criterion*.

Claim (2) is also difficult to defend: imagine that someone knew that, in the future, it will exist a person R-related to her. It is not implausible to believe that it would matter very little to her whether this R-related future person has the very same brain that she has now. It would not matter if the future person has a duplicate of her brain. What would matter to her would be the various relations between herself "and others, whom and what [she] love[s], [her] ambitions, achievements, commitments, emotions, memories, and several other psychological features."[330]

The person we are considering knows she will have a Replica of herself in the future, similar enough not to loosen significantly psychological connectedness (for example, the Replica has the same sex) and psychologically continuous with the herself that exists now. This Replica will appear when this person's original body is destroyed. This Replica would be like the previous person in all aspects that matter, but the psychological continuity would not be maintained with the normal cause, which is the maintenance of the same brain through time, and there would be no physical continuity. If this person considers not this Replica herself, she would believe that

(a) the replica is not herself

Which would consist in the fact that

(b) there is not physical continuity between her and the Replica

And in the fact that

(c) because of (b), R will not have the normal cause.

Since (a) consists in (b) and (c), it can be ignored and we should focus on (b) and (c). According to Parfit, both the facts cannot matter much. It can be stated that, though

[329] (Parfit, 1984, p. 284)
[330] (Parfit, 1984, p. 284)

the Replica is psychologically continuous with the former person, the Replica will not be the former person because of (b) and (c). "But this is not a further difference in what happens, beyond the difference in the cause".[331] It must certainly be admitted that the cause is not normal. Still, there is no motive for believing that the cause matters more than the effect. The effect is the same both in ordinary survival and in the creation of a Replica instead of the original body: the holding of Relation R. It is not rational to believe that it matters more the method of causation than the effect, if the abnormal method of causation has no different effect than the normal method of causation.

The only motive for preferring one's present brain and body to a non-coexistent, psychologically connected perfect Replica can only be a sentimental value, "like one's wish to keep the same wedding ring, rather than a new ring that is exactly similar."[332]

Therefore, Physical Continuity is *not* the criterion for Personal Identity. We do not feel our identity vanishing in the case of an exact Replica psychologically connected with us.

Also, it is not true that Personal Identity is Psychological Connectedness and/or Continuity with normal causes. Recall that the normal cause of psychological continuity involves the continued existence of enough of the brain. Since our identity is preserved with an abnormal case like the Replica, a normal cause is not necessary for preserving our identity.

For deciding if we need to accept Parfit's view according to which Personal Identity is Psychological Connectedness and/or Continuity with any cause, we need to reject the view according to which Personal Identity is Psychological Connectedness and/or Continuity with only reliable causes. A reliable cause of Relation R is any cause on which we can rely. For example, as already stated, the normal cause for sight is the naked eye; a reliable cause for sight is an eye helped by the lens of glasses.

"There is an obvious reason for preferring, in advance, that the cause will be reliable."[333] Suppose that someone, aiming to produce a teleporting device,[334] invents

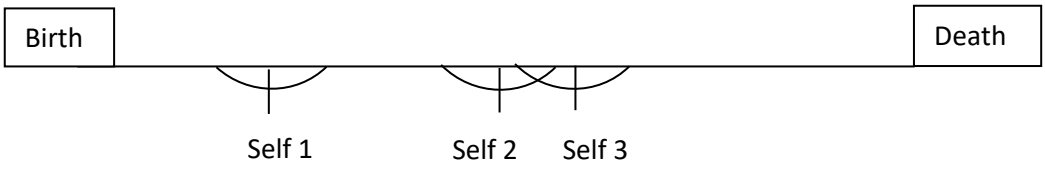
[331] (Parfit, 1984, p. 286)
[332] Ibid.
[333] Ibid.

an object that makes Replicas of individuals, not in the place in which the individual is, but somewhere else. But this inventor commits some errors in designing his device. With this teleporting device, the replacements of a person with her Replica would work perfectly in a few cases, but in most cases they would be a complete failure. In a few cases, the Replica would be identical to the former person. But in most cases the Replica would be totally unlike the former person. It would be irrational to ever use the device.

> But this is irrelevant. We should ask, 'In the few cases, where my Replica will be fully R-related to me, would it matter that R did not have a reliable cause?' I believe that the answer should again be No. Suppose that there is an unreliable treatment for some disease. In most cases the treatment achieves nothing. But in a few cases it provides a complete cure. In these cases, only the effect matters. This effect is just as good, even though its cause was unreliable. We should claim the same about Relation R.[335]

Therefore, the reliable criterion for Personal Identity is the *Widest Psychological Criterion*, according to which (1) There is psychological continuity if and only if there are overlapping chains of strong connectedness. *X* today is one and the same person as *Y* at some past time if and only if (2) *X* is psychologically continuous with *Y*, (3) this continuity has *any* kind of cause, and (4) it has not taken a 'branching' form. (5) Personal identity over time just consists in the holding of facts like (2) to (4).

We saw that personal identity is defined according to the *Widest Psychological Criterion*. We saw how our identity can be undetermined. We saw how our identity is not what matters, but rather, in our concern about our own future, what matters is Relation R, with any cause. We are close to state that Egoism is irrational.



[334] Parfit uses the example of such as kind of teleport much more frequently than I do. See for example (Parfit, 1984, p. 199-201, 279-280, 284-285 and others)
[335] (Parfit, 1984, p. 287)

Above there is a representation of the average life of a human being. The straight line that goes from her birth to her death represents Psychological Continuity. Usually a human being does not take a branching form, thus there are no forks in this line. A human being is usually psychologically continuous within her whole lifetime. But a human being is not psychologically *connected* within her whole lifetime. The segments called Self 1, Self 2 and Self 3 are three moments during the life of the average individual. In each of those three moments, the elements composing the Psychological Connectedness are different. The arches around the segments are all the moments with which each Self is connected. For example, Self 2 is connected to some extent with Self 3, but is not connected at all with Self 1. This means that the memory, the character, the desires, the affections, the knowledge and so on of Self 1 have something in common with Self 3, but nothing in common with Self 1. We can imagine Self 2 as who I am now, Self 3 as who I will be within three months and Self 1 as who was me when I was 6 years old. Self 2 is connected almost entirely with the Self that will succeed him two second later, is connected to some degree with Self 3, but has no connection with Self 1.

We have different psychological features in different moments of our lives. Within our life, there is a continue alternation of successive selves.

We have now all required for rejecting Egoism, or S. The central claim of S is the *Principle of Prudence*. According to it, the moral agent should have *impartial* concern for *all* parts of his or her conscious life. We now know that what fundamentally matters is Relation R, psychological continuity and connectedness, with any cause. *Both* relations matter. It is not defensible to claim that only continuity matters. As I stated in 7.1, suppose that in an hour I shall suddenly lose all my memory, or that all my present desires and intentions will be erased, or that all my affections will disappear, or that my character will not be recognisable. It would be paradoxical to state that anyone of those changes would not matter, and that I would not have any reason to avoid such a change, if I can. Since psychological connectedness matters, Parfit claims:

> My concern for my future may correspond to the degree of connectedness between me now and myself in the future. Connectedness is one of the two relations that give me reasons to be

> specially concerned about my own future. It can be rational to care less, when one of the grounds for caring will hold to a lesser degree. Since connectedness is nearly always weaker over longer periods, I can rationally care less about my further future.[336]

This cannot be defensively denied. If this is true, then the *Principle of Prudence* has to be rejected, and thus all S can be rejected. S might reply something like: "it is not true. All the parts of a person's future are equally part of his future". This is trivial truth, as it is a trivial truth the fact that "all of a person's relatives are equally his or her relatives". Be a person's relative is a matter of degree, as being a part of a person's future life is a matter of degree. In some sense, echidnas are my parents, since we have a common ancestor for reasons of biological evolution. My degree of kinship is nevertheless stronger with my siblings rather than with any kind of echidna. The same is true for being a part of someone's future. I was a part of my 6-years-old-self's future, but my psychological connectedness with him is close to none. This is how, if Parfit is right, the Dualism of Practical Reason is solved. Let us recap shortly the whole history of the Dualism.

In the first part of this work, Sidgwick showed the features of Intuitionism, Utilitarianism and Egoism. Sidgwick managed to demonstrate that Intuitionism and Utilitarianism works well only if unified, but he did not manage to resolve the contrast between this unified theory and Egoism. He called this problem Dualism of Practical Reasons, and believed it to be the profoundest problem in Ethics.

After him, Nagel tried to solve the Dualism by examining the concept of Reason, but he failed, since he was not able to find a concept of Personal Identity that admitted impersonal reasons as motivating for action.

Finally, Parfit tried to solve the Dualism. He first challenged Egoism with objections both from Consequentialism and the fictional Present-Aim Theory: since no theory of the three managed to defeat another theory with arguments, each had to qualify its claim. The central claim of Egoism is that it is supremely rational for an agent to benefit only herself, and have equal concern for all moments of her life. Through a metaphysical examination on Personal Identity, Parfit demonstrated that it is not rational for an agent to have equal concern for every moment of her life: who performs an action cannot benefit herself in the future, because in anyone's life there

[336] (Parfit, 1984, p. 313)

is a succession of selves. In fact, according to Parfit, Personal Identity consists in psychological connectedness and continuity, and a person preserves during her lifetime only continuity. Since what matters is connectedness, and connectedness changes, we cannot benefit ourselves in the future. Nor it would be supremely rational to benefit someone we are psychologically continuous with, since psychological continuity does not matter.

Since this is so, Egoism, or S, cannot qualify the claim according to which it is supremely rational for an agent to benefit only oneself, and have equal concern for all moments of her life. Therefore, Egoism has to be discharged: Parfit solved Dualism of Practical Reason in favor of Consequentialism.

Does the defeat of S imply the win of P rather than C? Does it mean that it is allowed to act according to P, and therefore disregard imprudence?

If we appeal to C, imprudence is immoral. C is agent-neutral and temporally neutral. It is wrong to damage another person, even if she is our future self. This is clearly not enough to reject P. But, in order to reject P, it is required to define which version of P is the best. It cannot be CPS, since its central claim, according to which it is supremely rational for someone to benefit herself equally trough time, has been rejected. It might be that a development of PC2, that was a general, unrefined theory of Altruism, is the best version of P. Parfit does not examine which is the best version of P, and neither it will be examined in this work.

Parfit will examine how the conception of Desert,[337] commitments[338] and distributive justice[339] has to be revised according to his concept of Personal Identity. As shown in 3.2, their definition and application was indefinite in Common Sense Morality. In 3.4 it was also shown how there were some discrepancies between Utilitarianism and contemporary legal justice. We will not see Parfit's answers about those problems.[340] We will neither see if this view about Personal Identity is a depressing or

[337] (Parfit, 1984, p. 323-326)
[338] (Parfit, 1984, p. 326-329)
[339] (Parfit, 1984, p. 329-330)
[340] Parfit solves another problem. In note 253 I wrote how Nagel objected that Utilitarianism and Consequentialism overlook the differences between individuals. Parfit answers to this objection too, in (Parfit, 1984, p. 330-332). I here only briefly suggest to the reader that, according to Parfit's analysis on Personal Identity, not only differences between individuals are considered, but also between the same individual at different times. It is right to state that his version of Consequentialism does not consider relevant difference between individuals, but it is wrong to state that a consequentialist considers all lives influenced by his act as unified in only one life, that is the one lived by the person that is evaluating the consequences.

unbelievable view about ourselves.[341] In this work, not even part Four of *Reason and Person*, which is seminal for Population Ethics, will be analysed in its fullness. But Population Ethics, briefly introduced in 4.1 when talking about *Average* and *Total Principle* and the *Repugnant Conclusion,* will not be disregarded. On the contrary, in the last part of this work, an application of Parfit's concept of Personal Identity to Population Ethics will be proposed.

[341] (Parfit, 1984, p. 274-280)

Part 3: Possible developments

Chapter 8: Revising Beneficence

At the end of the first part of this work, in 4.2, some problems of Utilitarianism has been listed. In the previous part I examined how Parfit resolved one of those problems, which is the Dualism of Practical Reason. In this part I will examine how another of the problems of Utilitarianism, namely the problem of the contrast between *Average* and *Total Principle*, might be solved. The field in which those principles are compared is Population Ethics. In what follows on Population Ethics, Parfit's concept of Personal Identity –according to which, as explained in chapter 7, identity consists in psychological connectedness and continuity – will be an extremely useful tool.

Before beginning our examination, let us define what Population Ethics is.

Population Ethics is the field of philosophical moral reasoning that seeks "an adequate theory of population value where [a] the number of people, [b] their welfare, and [c] their identities may vary".[342]

Let us explain the terms used into this definition. A population is defined as a set of lives. The welfare of a life is how well the life is going for who lives it, or how much this person enjoys of what makes life worth living. The higher the welfare enjoyed in someone's life, the better the life of that person. In Population Ethics it is commonly assumed that welfare can be negative, and therefore lives can be not worth living. The sentence "a person A's life is better off than a person B's" can be restated as "person A's life has a higher welfare than person B's life".[343]

I will not explain now the meaning of identity and its variation in Population Ethics since everything about this point will be examined in 8.1 and following.

In order to find a theory for population value it seems necessary to define a Population Axiology. A Population Axiology aims at establishing the criteria according to which we ought to prefer a population rather than another; a Population

[342] (Arrhenius & Campbell, Forthcoming). I added the letters [a],[b] and [c], since they will be useful in 8.3.

[343] This is Arrhenius' terminology. See (Arrhenius, 2000, p. 6-12) for further explanations on the choice of this term.

Axiology should be able to rank every possible population from the most preferable to the least appealing.

It has been stated in this work that Part 4 of *Reasons and Persons* is seminal for Population Ethics. In Part 4 Parfit's main concern is not actually understanding how a moral agent ought to influence the number of future people and allocate welfare. His aim is rather finding a "*Theory X*", which is a revision of the Principle of Beneficence,[344] or, as Parfit puts it, "the best account of Beneficence".[345] The Principle of Beneficence is the principle according to which we should try to do what would benefit people most, at least sometimes. This principle seems uncontroversial, and is accepted by nearly all moral theorists. Unfortunately, this Principle is unclear in suggesting how to behave in certain specific cases, called by Parfit *Different Number Choices*. *Different Number Choices* are choices in which every alternative leads to the existence of a different number of people. Examples of *Different Number Choices* will be present in 8.1 and following. The aim of Population Ethics, which is finding "an adequate theory of population value where [a] the number of people, their [b] welfare, and their [c] identities may vary", can be restated as "solving *Different Number Choices*", or "finding *Theory X*".[346] *Total* and *Average Principle* are two proposals of *Theory X*, and we need to find which one of them, if any, is fit for solving *Different Number Choices*.

In 8.1 it will be explained what is intended by Parfit for *Different Number Choices* and consequently it will be clarified what Population Ethics is about. In 8.2 it will be made a remark about *Same Number Choices*, a kind of choice that affects the identity of future individuals but not their number, as I'll explain in 8.1. It will then be claimed, in 8.3, that *Different Number Choices* and *Same Number Choices* present alike difficulties, and they should consequently be treated alike. If this claim is right, the results of Population Ethics should be extended to some *Same Number*

[344] (Parfit, 1984, p. 366). Parfit, in (Parfit, 1984, p. 69) ascribes the *Principle of Beneficence* to Sir William David Ross (1877–1971), moral philosopher and commentator of Aristotle. One of his greatest contributions to Ethics is the statement of moral principles, such as the *Principle of Beneficence*, that a moral agent is *not* obliged to follow, but has *prima facie* reasons to follow. His most important work in moral philosophy is considered *The Right and the Good,* first printed in 1930.

[345] It is stated several times in (Parfit, Reasons and Persons, 1984). See for example (Parfit, 1984, p. 401)

[346] Notice that Population Ethics is not necessarily a field that concerns merely Consequentialists. It concerns all thinkers who believe that benefiting someone is a moral action and damaging her is immoral, since Population Ethics seeks how we should benefit future generations.

Choices.

If this extension is legitimate and the conclusions of Population Ethics can be applied to *Same Number Choices*, there will be consequences not only for the applicability of Population Ethics over moral questions concerning individuals' lives, but also for what criteria a *Theory X* should respect for being acceptable. These consequences will be explored in chapter 9, where it will be argued that the *Total Principle* is probably *Theory X*. In chapter 10 it will be explored how reasoning probabilistically may help in Population Ethics.

8.1. Three kinds of choices

Some of Parfit's claims about Personal Identity may have been hard to believe. However, the following assumption is not so hard to believe. Parfit states:

> *The Time-Dependence Claim*: If any particular person had not been conceived when he was in fact conceived, it is in fact true that he would never have existed.[347]

This claim is plausible since, if the conception of someone happens sooner or later than when she was in fact conceived, the cells involved in the birth would have been different, and therefore the person would have been different.[348] A similar, uncontroversial claim is that, if the two persons that conceived a child would have not conceived in fact a child (for example because they never met), this child would have never existed.

Some of our choices may affect the time in which a person is conceived. Some other may affect who meets whom. This means that our choices can affect the identity of future people. For example, as Parfit remarks,[349] the choice of building railways and motor cars changed the possibility of meeting people, and consequently the existence of future people.

Furthermore, as previously written, we can affect the number of existing people. An immediate example is a couple that decides to have a child, adding a future person.

---

[347] (Parfit, 1984, p. 351)
[348] Some objections and some alternative views are possible. I believe Parfit answers to the most relevant of them in (Parfit, 1984, p. 351-355)
[349] (Parfit, 1984, p. 361)

We can affect both the identity and the number of future people. This means that we can make three kinds of choices.

Same People Choices are choices whose consequences do not affect neither the number nor the identity of future people. According to Parfit, most of our moral thinking is about such choices, but "such choices are not as numerous as most of us assumes".[350] In 8.2 it will be argued that *almost no morally relevant choice is a Same People Choice*.

Same Number, Different People Choices (hence *Same Number Choices*) are choices whose consequences affect the identity of future people, but not their number. A couple that decides to postpone the conception of a child because it cannot afford it is an obvious example. There is another, more common but less obvious example, that will be an important subject in 8.2.

Different Number, Different People Choices (hence *Different Number Choices*) are choices whose consequences affect both identity and number of future people. This kind of choices is the main interest of Population Ethics. In 8.3 one of my aim is to show how the difference between *Same Number* and *Different Number Choices* is less definite than it seems.

Same People Choices will not be examined much because, as it will be shown in 8.2, most of them are trivial and because the Consequentialist Theory so far proposed gives enough guidance to suggest to a moral agent how she ought to act when facing those choices: she ought to choose what maximizes the good.

The Consequentialist Theory might give guidance also in the other two kinds of choices, but first the usage of the following, apparently self-evident principle has to be restricted in its usage, due to its unfitness in *Same Number* and *Different Number Choices*:

> *The Person-Affecting View*, or *V*: It will be worse if people are affected for the worse.[351]

Consider, for example, the present climate change. Suppose that all theories according to which humans have no responsibility for this kind of changes have to be

[350] (Parfit, 1984, p. 356)
[351] (Parfit, 1984, p. 370)

rejected and that the problem of climate change can be faced in only two ways. We can (a) continue with the present lifestyle, or insist even more in policies unhealthy for the planet, ignoring the fact that pollution worsens the Earth's condition. In order to avoid any kind of suffering caused by our lifestyle, we agree in avoiding to have children by inventing and using an infallible contraceptive, so that no one will live in a planet unfit for habitation. Alternatively, we can (b) find a policy that saves our planet, or find another planet in which future human beings will be able to live a life worth to.

Many will find (b) the right policy, but it must be noticed that (a) does not harm anybody. Nobody is harmed in (a) because, when planet Earth will not be fit for life, there will not exist any future person that will be affected for the worse, since due to the contraceptive no one will live in the future. If no one lives, no one is affected for the worse.

In (a) nobody is affected for the worse, therefore, if *V* is not rejected, (a) is a choice as good as (b). Maybe (a) is even better: in fact, in (b) people might be required to give up an important part of their wellbeing in order to save the planet, so those people would be affected for the worse. It is hence possible that, if *V* is not rejected, (a) has to be preferred to (b), because in (a) neither present people, that maintain the previous quality of life, nor future people, which do not exist, are affected for the worse, whereas in (b) present people is damaged for the worse. Many would find absurd to state that (a) is as good as (b), or maybe even better. Those people find that the *Person-Affecting View*, or *V*, is unfit for performing the choice between (a) and (b).

The following chart recaps who is affected for the worse in each choice.

	Choice (a)	Choice (b)
Is the present population affected for the worse?	No, since it does not change its lifestyle.	Maybe yes, since this population might be required to sacrifice a part of its wellbeing.
Is the future population affected for the worse?	No, since it does not exist.	No.[352]

As stated, choice (a) does not affect anyone for the worse.

Still, many accept intuitively (b) rather than (a), but it is not clear what principle they rely on. This is so because climate change policies are *Different Number Choices*. For solving those choices the best account for beneficence, or *Theory X*, is required. Unfortunately, it is not intuitively clear what this theory should be.

Let us see now what is usually considered a simpler kind of choice in which *V* is unreliable. The following is Parfit's example of a *Same Number Choice*.[353]

Consider a 14-year-old girl who wants to have a baby. Let us call this future baby John. The girl is very young, so it is very likely that John will have a bad start in life, which will affect his adult life. Nevertheless, John's life will predictably be worth living. If the girl waits some years before conceiving, she would have a different child, called for example Jack, to whom she would be able to give a better start in life.

We might try to warn the girl by pointing out how generating John will be worse for her, but she would reply that this is her affair, and that she has the right to do what she wants, unless her action is bad for someone else. We should therefore show how her act damages someone else: the child. But we can't. If we admit *V* we have no ground for convincing this girl not to have John. Since John's life will

[352] No, unless causing someone to exists thereby damages this person. The two most accepted views are actually that causing someone to exist benefits thereby this person or does not damage nor benefit this person. This debate will not be considered in this work, since it is not necessary. For some information about this matter see (Parfit, 1984, p. 487-490).

[353] This example is taken from (Parfit, 1984, p. 358), with slight changes.

predictably be worth living, if the girl has a baby now she will harm no one. She in fact cannot harm Jack, since Jack would not ever exist.

The following chart recaps who is affected for the worse in each choice.

	Generating John	Generating jack
Is John affected for the worse?	No, since he will predictably have a life worth living.	No, since he will not exist.
Is Jack affected for the worse?	No, since he will not exist.	No, since he will predictably have a life worth living.

Neither choice affects anyone for the worse.

To recap, *The Person-Affecting View*, or *V*, has to be considered unfit for *Different Number* and *Same Number Choices*. This is so because there are choices that are bad even if their consequences affect no one for the worse. In the case of climate changes, if every present person uses the contraceptive, there will not exist a generation suffering for climatic changes, therefore no one will be affected for the worse; still, it seems morally better if present people preserve the planet's inhabitability. In the case of the 14-years-old girl conceiving John will not harm him, since his life will be predictably worth living, neither will harm Jack, since he would not exist. Still, the choice of waiting the conception seems morally better.

Parfit states that, in order to convince the girl to wait before conceiving a child, we must not rely on *V,* but rather on

> *The Same Number Quality Claim*, or *Q*: If in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have a lower quality of life, than those who would have lived.[354]

[354] (Parfit, 1984, p. 360). Someone might think that, instead of *Q*, we might appeal to some system of rights, or to some contractarian view. We cannot, since it is hard to believe that people that will never exist have rights, or might agree to a contract. For discussions on this matter see (Parfit, 1984, p. 364-366, 391-393,490-493).

Parfit claims that *V* is fit for *Same People Choices*, but unfit for *Same Number Choices*. He claims that *Q* is fit both for *Same People* and for *Same Number Choices*,[355] because in *Same People Choices Q* coincides in fact with *V*. *Q* is a general principle, of which *V* is a particular specification in a small number of occasions.

For *Different Number Choices* we need a principle more general than *Q*. Parfit calls this principle *Theory X*. The aim of Population Ethics can be considered finding a satisfactory *Theory X*,[356] a theory reliable in *Different Number Choices*, that coincides with *Q* in *Same Number Choices* and coincides with *V* in *Same People Choices*. The fact that the results of Population Ethics, conceived as proposals of plausible candidates for *Theory X*, should be acceptable also on *Same Number Choices* has been often overlooked by thinkers of Population Ethics, whose possible application on individual's lives has been therefore underexplored.

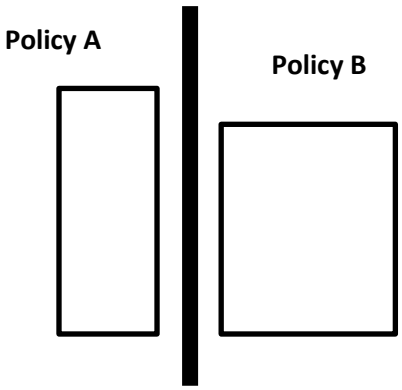
It is now the case to show why *Q* cannot help in *Different Number Choices*. Before that, some assumptions are needed.

Assume, as Sidgwick,[357] that lives can be worth living. Assume that a person's life can be better or worse off than another's life. Assume that someone's life's welfare can be measured and compared with someone else's life's welfare.[358] Remember that the welfare of a person's life is, roughly, how well this person is living. Assume that a life is considered having welfare 0 if and only if it has the following feature: if someone could choose between living or not living such a life, she would be indifferent.[359] Assume that it is possible to have a life that has welfare higher or lower than 0. Welfare will be not defined precisely, since it is enough to

[355] As I will make clear in 8.3, I think *Q* would be fit only in a minority of the *Same Number Choices*.
[356] As I said, *Theory X* is a revision of the *Principle of Beneficence*, according to which we should try to do what would benefit people most, at least sometimes. Therefore *Theory X* should be our best and most complete account on beneficence. I wrote that in (Parfit, Reasons and Persons, 1984) *Theory X* is a theory that must help in *Different Number Choices*. This is true only until Parfit mentions *the Repugnant Conclusion*. From that moment onward, in (Parfit, Reasons and Persons, 1984) *Theory X* becomes a theory that (1) must help in *Different Number Choices* and (2) avoids the *Repugnant Conclusion*. Since it is doubtful whether (2) has to be accepted in our best and most complete account of beneficence, I shall consider *Theory X* without feature (2). I write that it is doubtful because arguments supporting *the Repugnant Conclusion* has been provided in (Huemer, 2008), (Tännsjö, 2004), (Ng ,1989) and several other articles. I myself believe that it has to be accepted, as it will result clear in 9.2.
[357] See 4.1
[358] This measurement and comparison needs not to be precise.
[359] For example, this life might be constantly in a situation like what Sidgwick called the Hedonistic Zero, as stated in 2.3.

know what follows: if a person *M* has a welfare higher than a person *N*'s welfare, *M* is better off than *N*. Finally assume, for simplicity, that lives in a population share the same welfare.

Suppose now that a policymaker has to make a choice between two policies. The outcomes of the two policies are shown below.



The width of the two rectangles represents the number of people (and consequently of lives) generated by the policy. The height represents their welfare.

All lives generated by Policy A are better off than lives generated by Policy B. Could this fact be morally outweighed by the fact that in Policy B there are significantly more people?

Q cannot answer to this question. *Q* does not state if the existence of more people makes a policy better than another. Since we cannot rely on *Q*, we will rely on a theory that coincides with *Q* in *Same Number Choices*[360] and with *V* in *Same People Choices*, and that helps in solving *Different Number Choices*. If we believe that the *Average Principle* is *Theory X*, we answer "No" to the question "does the fact that in Policy B there are significantly more people morally outweigh the fact that all people generated by Policy A are better off than people generated by Policy B?". If we believe that the *Total Principle* is *Theory X*, we answer "Yes". Those principles have been introduced in 5.1. Let us restate them now, with their formulas. The word "impersonal" will be added to those principle because, in stating them, we cannot rely on the *Person-Affecting View*, or *V*, because it is not fit for *Same Number* and *Different Number Choices*.

[360] Or at least some *Same Number Choices*, as I will point out from 8.2 onwards.

> *Impersonal Average Principle*: If other things are equal, the best outcome is the one in which people's lives go, on average, best.[361]

If we call *N* the number of existing people, individual 1, individual 2,individual 3… individual n are called $p_1,p_2,p_3...p_n$, and $W(p_x)$ is a function which takes individuals as arguments and gives their welfare[362] as values, the *Average Principle* would prescribe that the best moral choice is the one that maximize the result of the following formula:

$$(W(p_1) + W(p_2) + W(p_3) \dots + W(p_n))/N$$

> *Impersonal Total Principle*: If other things are equal, the best outcome is the one in which there would be the greatest quantity of good.[363]

According to the *Impersonal Total Principle*, the quantity to maximize is expressed by the following formula:

$$(W(p_1) + W(p_2) + W(p_3) \dots + W(p_n))$$

In 4.1 it has been showed how the *Impersonal Average Principle* leads to morally absurd choices. Gustaf Arrhenius synthesized them stating that the *Impersonal Average Principle* violates the following, extremely plausible principle:

> *The Non-Sadism Condition*: an addition[364] of any number of people with positive welfare is at least as good as an addition of any number of people with negative welfare, other things being equal.[365]

---

[361] (Parfit, 1984, p. 386).
[362] In 4.1 "welfare" was replaced with "happiness" since we were considering Sidgwick's Utilitarianism.
[363] (Parfit, 1984, p. 387)
[364] I remember that the expression "adding people" means "causing to exist more future people".
[365] (Arrhenius, 2000, p. 64, 203)

I showed how the *Non-Sadism Condition* is violated by the Average Principle in 4.1. For sake of clarity, I do it again here. Suppose that we live in a world with a population of 8 billion people, roughly as much as today, where everyone lives a really good life. Imagine that we are obliged to add people, and that we have this only alternative: either we can choose (1) a policy that causes the existence of 4 billion people whose life will be good, but less good than the life of the already existing 8 billion people, whose lives are really good lives; or we can choose (2) a policy that causes the existence of a single person, whose life will not be worth living, and full of suffering rather than happiness. According to the *Average Principle*, we ought to choose (2), preferring the existence of a suffering person than the existence of happy people.[366]

The Non-Sadism Condition is part of a list of conditions that, according to Arrhenius, any satisfactory Population Axiology (or any *Theory X*) should not disregard in order to be acceptable.[367] *The Non-Sadism Condition* is very plausible. To my knowledge, all thinkers of Population Ethics agree with Arrhenius in stating that a proposal of *Theory X* cannot violate this condition.

On the other hand, the *Impersonal Total Principle* implied[368]

---

[366] I can restate what just said with numbers. Suppose that the 8 billion people, that we call group A, have welfare 4 *pro capite*, while the 4 billion people, or group B, have welfare 2 *pro capite*, and finally the suffering person, or group C has welfare -8. (A negative welfare implies a life not worth living). The average welfare of (1) would be the average of groups A and B. The total welfare in A is $(8 \times 10^9) \times 4 = 32 \times 10^9$, the total welfare in B is $(-4 \times 10^9) \times 2 = 8 \times 10^9$, the average is the sum of the two total welfares divided by 12 billion people, that is the total of persons living in group A and B: the average welfare of (1) is thus $(32 \times 10^9 + 8 \times 10^9) / (12 \times 10^9) = 3,(3)$. On the other hand, the average welfare of (2) would be the average of the groups A and C. The total welfare in A is always $(8 \times 10^9) \times 4 = 32 \times 10^9$, the total welfare in C is $1 \times -8 = -8$, the average is the sum of the two total welfares divided by 8 billion people and one: the average welfare of (2) is therefore $(32 \times 10^9 + (-8)) / (8 \times 10^9 + 1) = 4$ circa. The -8 of population C is trivial when counting the average. Since 4 is bigger than 3,(3), according to the Average Principle (2) is to be preferred to (1).

[367] The complete list of condition can be found in (Arrhenius, 2000, p. 202-204)

[368] I repeat here what said in 4.1, note 209, in order to demonstrate this implication. In what follows N is the number of existing people; $p_1,p_2,p_3...p_n$ stand for individual 1, individual 2,individual 3… individual n; $W(p_x)$ is welfare enjoyed by an individual x; U is the sum of $W(p_1) + W(p_2) + W(p_3) ... + W(p^n)$. Remember that the *Total Principle* aims at maximize *U*. Let us assume that welfare can be quantified, and that a person that lives a life barely worth living enjoys a welfare of 1. Imagine any possible population of any value *N* and in which everyone lives a life that is more than barely worth living, and thus the value of *U*/*N* (average welfare) is greater than 1. Imagine now a possible population of *N'* people, where the number of people *N'* is greater than the *U* of the other population. Imagine now that everyone in this population of *N'* people has a life that is barely worth living (that has welfare 1). Since each person has only welfare 1, the value of *N'* is equal to the value of the total amount of happiness *U'* of that population. By hypothesis, *N'*>*U*, but since *N'*=*U'* we can conclude that *U'*>*U*. Therefore, the population in which everyone has a life barely worth living results better, *quod erat demonstrandum*.

> *The Repugnant Conclusion*: For any possible population of [. . .][369] people, all with a very high quality of life,[370] there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living.[371]

Many thinkers of Population Ethics, such as Derek Parfit, tried to find a *Theory X* more satisfying than what proposed by means of the *Average* and the *Total Principles*. Others[372] accepted *the Repugnant Conclusion*. Others, such as Gustaf Arrhenius, believed that a satisfactory *Theory X* that is effective in *Different Number Choices* cannot be found, and proved their belief with formal arguments.

It will be argued in 9 that, if we consider more carefully *Same Number Choices*, *the Repugnant Conclusion* receives some support. It will also be argued that, if extending Population Ethics to individual choices is legitimate and thinkers who believe that it cannot be found a satisfactory Population Axiology (and thus a satisfactory *Theory X*) are right, morality's trustworthiness is under great threat.

8.2. Choosing who lives our life

Return now to *Same Number Choices*. Those choices influenced the identity, but not the number, of future people. I believe that a very relevant kind of those choices has been neglected. Consider

> *Alexander's choice*. In high school Alexander was very good in chemistry, and developed an interest for medicine. The first day after having finished high school, he receives an appealing job offer for working abroad in a platform that extracts natural gas. He still considers the possibility of becoming a physician. He is free from any constraint in choosing between accepting the job offer or going to university and studying medicine.

---

[369] In the original formulation, Parfit writes "any possible population of *at least ten billion people*". I omit the italic part. He chooses the number of ten billion people because it is a number similar to the present number of existing people, as stated in (Parfit, 1984, p. 402-404): he believed that a similarity of number in the present state of things and in *the Repugnant Conclusion* would have made this latter simpler to evaluate.

[370] Parfit's expression "quality of life" is equivalent to what is here called "welfare".

[371] (Parfit, 1984, p. 388)

[372] See note 356.

Suppose now that twenty years have passed from the moment in which Alexander has made his choice. He might have become a physician or a worker on the platform. If twenty years later he is a physician, he might have strengthened friendships in the hospital in which he works, he might have forgotten a part of his knowledge in chemistry while his proficiency in medicine has been enhanced, he might have kept in touch with his family, high school friends and so on.

If instead twenty years later he works on the platform, he will have strong friendships with people on the platform, he might have lost passion in medicine and improved his proficiency in chemistry, he might be less in touch with his former family and friends, and so on.

In the two possible futures, Alexander develops different interests, feels affection for different persons and remembers different things. If Parfit is right, Personal Identity consists in psychological continuity and connectedness, with any cause, as showed in 7. Being psychologically connected with a former self means to share most of its desires, passions, affections, memory, ideals and so on. Being psychologically continuous with a former self means holding with it overlapping chains of connectedness.

Alexander the Physician is psychologically continuous with the Alexander of 20 years before, but he is not psychologically connected with him. The same can be said of Alexander on the Platform. This is nothing new.

What is relevant here is that Alexander the Physician is *not* psychologically continuous *nor* psychologically connected with Alexander on the Platform, and *vice versa*. They are not connected since they share little to none common psychological features, and they are not continuous since overlapping chains of strong connectedness between them do not exist. They are mutually exclusive future selves. Only one will exist. This means that Alexander, when choosing his career, is making a *Same Number Choice* as much as the choice of the 14-years-old girl is a *Same Number Choice*. Alexander chooses who exists and who does not. The only difference from the 14-years-old girl is that the life of Alexander is not caused to exist by this choice, but the possibility of a future Alexander rather than the other is realized by this choice. Alexander's choice is a *Different People, Same Number Choice* that is realized within a single life, while the choice of the 14-years-old girl

concerns two different lives, John's and Jack's lives. Alexander is choosing who will live his future life, the 14-years-old girl is choosing who lives a life.

Choices such as Alexander's are not uncommon. Everyone makes choices about their friendships, loves, hobbies, the places where they spend time, their career, their readings. *Maintaining* a friendship, a love, a hobby etc. etc. can be considered a choice as well, even if who is choosing is not always completely aware of them. All those choices, that are made *daily*, influence our future selves. Almost every day we make *Different People, Same Number Choices*, which means that almost every day we affect the existence of a future being.[373]

Does this imply that *Same People Choices* do not exist? No, they do exist, but they are few. They are (1) some of the choices concerning a person that is about to die,[374] (2) morally trivial choices or (3) both (1) and (2).

Choices of the set (1) are *Same People Choices* because, when the one who will feel the effect of the choice is close to death, it is possible that no future self will be affected by the choice. Imagine, for example, that someone is in agony, and is going to die within a few minutes; an agent might choose to kill him, in order to shorten the dying person's suffering. This would be a *Same People, Same Number Choice*. *V* would be fit for it.

An example of (2) is the following. Choosing to eat pasta over rice for a day will not plausibly affect who will be my future self. But preferring pasta over rice is not a morally relevant choice, at least in the greatest majority of cases. *Any* choice that affects the identity of future selves is a *Different People, Same Number Choice*. It is hard to think of a morally relevant choice that does not affect any future self except some cases concerning a person close to death: all other cases of morally relevant choices I am able to conceive affect the future life of the agent or someone else's life. It might even be claimed that a choice that does not concern a dying person is as morally relevant as it changes someone's future self (not necessarily the agent's future self), but a decisive arguments for showing that this is

[373] Clearly, affecting the existence of a future being implies a moral responsibility, and therefore those kind of *Same Number Choices* implies a moral responsibility as much as the choice of the 14-year-old girl implies moral responsibility.

[374] Only some cases. In fact, if an agent benefits someone else, and the benefited person is about to die, it is possible that the kind act of the agent will have some influence in the *agent's* future selves.

certainly true seems to lack. What is hard to deny is that any choice that certainly changes someone's future self is a morally relevant choice.[375]

It is important to remember that, as shown in 8.1, *V* should be rejected in all choices that are not *Same People Choices*. Since all choices that are not *Same People Choices* are the majority of relevant choices, or maybe all relevant choices that do not concern dying persons, *V* should be considered unfit for the greatest majority of the morally relevant choices.[376]

It has been argued that *Same People Choices* are less relevant than *Different People, Same Number Choices*. It will be argued now that *Same Number Choices* often raise difficulties much like *Different Number Choices*. Those difficulties are so alike that a theory fit for helping in a kind of choice is very likely to help also in the other; consequently, a theory unfit for a kind of choice is unfit also in the other kind of choice. *The Same Number Quality Claim*, or *Q*, will be shown unfit for some *Same Number Choices* as much as it is unfit for *Different Number Choices.*

8.3. Lifespans

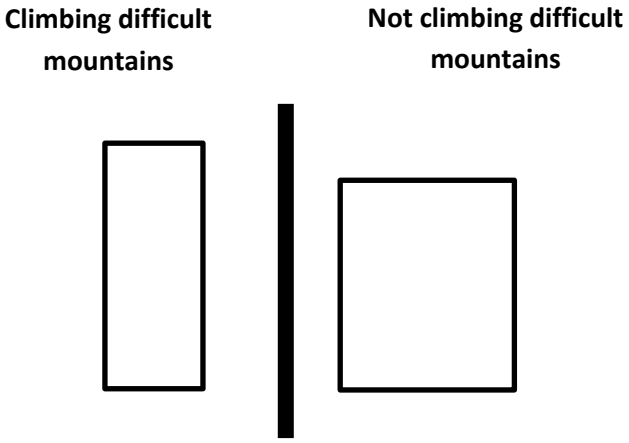
Imagine

> *The Climber's choice*. A person finds out that she loves climbing mountains. This hobby becomes very important for her. The more the mountain is difficult, the more she enjoys climbing it. She is aware that, unfortunately, difficult mountains have a higher chance to kill her. She is wondering if she should avoid difficult mountains, in order to lower the risk to get killed. This would make her hobby less enjoyable.

[375] It might be objected that *Alexander's Choice* is not a moral choice, but rather an existential choice, since the choice of a career seems not to matter morally. It is an existential choice only if Alexander ignores how his choice will influence his lifespan and his welfare. In this case, due to Alexander's ignorance, the alternatives seems to lead approximately to the same length of the lifespan and welfare enjoyed. Imagine on the opposite that Alexander knew that, for example, the career of a physician will lead him to have a shorter life and to have a lower welfare than the alternative. Choosing the career of the physician would be defined imprudent by someone. If someone accepts Parfit's theory of Personal Identity according to which identity does not matter, she would not call it imprudence, but rather immorality, because the choice damages a future being. Further on Responsibility will be stated on 9.2. What matters now is that Alexander's choice seems not a moral choice only as far as we ignore how to evaluate it from a moral point of view, that is to say as far as we do not know how the welfare and the length of his life will be affected by the choice.

[376] Parfit wrote in (Parfit, 1984, p. 370) that, even if we discover *Theory X*, we might appeal often to *V* because it is more familiar to us. It is possible that we might do so, but we would be wrong. If what has been written in 8.2 is right, we should avoid *V* in the majority of relevant moral choices.

Should the Climber make this choice? The two blocks below this paragraph represents the most likeable outcomes of her choices. The width of the squares is the length of her life, the height of the squares represents her welfare.



In tracing those blocks it has been assumed that a person's life can be better off or worse off if she makes a choice rather than another. It has been assumed that a life is considered having welfare 0 if and only if it has the following feature: if someone could choose between living or not living such a life, she would be indifferent. It is assumed that it is possible for a life to have a welfare higher or lower than 0. It has also been assumed that alternative welfares can be measured and compared one another. For simplicity, let us assume that, for the climber, the only thing that makes a life worth living (that is to say, the only things that makes welfare positive, negative, high, low or zero) is climbing mountains. In this way we can ignore all other events that may condition the life of the climber, in order to ignore in the blocks, for example, the fact that dying falling from a mountain is likely to be extremely painful. Always for simplicity, let us assume that every moment of the Climber's lifetime will have the same welfare. It is also assumed the correctness of what written in 8.2 about future identities, which is that any choice that affects the future identity of an existing person is *not* a *Same People Choice* (unless it is a morally trivial choice or it concerns a dying person) and therefore those choices cannot be made relying on *V*.

All moments of the life generated by the choice of climbing difficult mountains are better off than moments of the life generated by the choice of *not*

climbing difficult mountains. Could this fact be morally outweighed by the fact that, if the climber does not climb difficult mountains, she lives longer?

Notice that the just stated question is extremely similar to the question asked in 8.1 when evaluating the *Different Number Choice* between a Policy A that caused an higher average welfare shared by less people's lives and a Policy B that caused an higher total welfare allocated among more people's lives. The question was: "All people's lives generated by Policy A are better off than people's lives generated by Policy B. Could this fact be morally outweighed by the fact that in Policy B there are significantly more people?". Furthermore, the blocks that represent *the Climber's choice* share the features of the blocks that represent the outcomes of the *Different Number Choice* between Policy A and Policy B in 8.1. The feature that is more clearly common in those representations is that the choice is between an outcome with more total welfare and an outcome with more average welfare. Those similarities are not a coincidence: they reveal that *Same Number* and *Different Number Choices* are much more alike than what has been so far usually believed by the majority of thinkers. I will now show a further difference: namely that, as *Q* is unfit for *Different Number Choices*, it is unfit for *Same Number Choices*. This leads to think that we require *Theory X* for solving not only *Different Number Choices*, but also some *Same Number Choices*.

It has been claimed in 8.2 that certain choices in the life of an individual affect the identity of her future selves. *The Climber's choice* seems a choice of this kind. Those kind of choices are traditionally called *Same Number Choices.* According to Parfit, *Same Number Choices* can be made relying on

> *The Same Number Quality Claim*, or *Q*: If in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have a lower quality of life, than those who would have lived.[377]

Call *CLife 1* the life of the climber if she chooses to climb difficult mountains and *CLife 2* the life of the climber if she does not choose to climb difficult mountains. Any moment of *CLife 1* has a higher welfare than *CLife 2*, but *CLife 2* leads to a longer lifespan than *CLife 1*. If we rely on *Q*, then it is unclear which choice is better,

[377] (Parfit, 1984, p. 360).

because *Q* does not state if living longer makes a moral difference. *Q* is unfit to solve those kind of problems exactly as it is unfit to solve problems raised in *Different Number Choices*. Once again, in order to decide between *CLife 1* and *CLife 2* we need the best account for beneficence: *Theory X*. Finding *Theory X* is the aim of Population Ethics. If we are Consequentialists, we might rely on an *Average Principle* believing it to be *Theory X*, and state that *CLife 1* is better off than *CLife 2;* we might instead believe that the *Total Principle* is *Theory X*, rely on it, and prefer *CLife 2* to *CLife 1*. We might also rely on a third principle we believe to be *Theory X*, but we certainly cannot rely on *Q*, exactly as in *Different Number Choices*.

Return now to *Alexander's Choice.* We have no way to know if this choice raises or not problems similar to the ones observed in *Different Number Choices*. It is difficult to state if becoming a physician will lead Alexander to live longer or shorter than working on the platform. Likewise, it is not clear whether becoming a physician will let Alexander enjoy a higher or lower level of welfare than working on the platform. Nevertheless, it is very likely that (1) Alexander the Physician and Alexander on the Platform will enjoy different things in their lives (hence they will probably experiment a different welfare), and it is unlikely that (2) they will have exactly the same lifespan. *Alexander's Choice* might have been a choice between a life with a higher total welfare and a life with a greater average welfare. It is possible that, if Alexander would have known how his lifespan would have been influenced by his choice, *Q* might have been unfit also for *Alexander's Choice*, and a *Theory X* might have been required instead. Since (1) is very likely and (2) is very unlikely it follows that, in *Alexander's Choice* and all choices alike, *Theory X* is required, if not in practice at least in theory. The *Same Number Choices* in which *Q* is enough would be, if we knew how the lifespan is influenced, only a minority, because it is unlikely that the alternatives lead to lives equally long.

We do not know, in cases like *Alexander's Choice*, how the lifespan will be probably affected. The cases in which this is not clear are much more frequent than cases like *The Climber's choice*. In *The Climber's choice* was clear, even if approximately, how the quality and duration of life is likely to be affected. In cases like *Alexander's Choice* it might be unclear whether we should rely on *Q* or we need a *Theory X*, because we cannot understand in which alternative the lifespan will be

shorter. If we knew the length of the possible lifespans, *Q* would be unfit in all *Same Number Choices* in which the lifespan is affected differently by the different alternatives - that is to say, in the greatest majority of the *Same Number Choices*. Instead of *Q* we would need a *Theory X*. Since we may not know how the alternatives affect the lifespans, or since, once we knew the difference in lifespans, this difference might be irrelevant, in our practical use of morality we may still rely on *Q*. This does not make *Q* a completely reliable principle, because it is fit only if we lack of an important information: the information concerning lifespan. If we knew the lifespans' length generated by the choices, *Q* would be unfit. A reliable theory should be more precise when more information is available. *Q* is not reliable: if whoever uses *Q* has more information (which is the information about lifespans) *Q* gives no answer. We need therefore a *Theory X* that coincides with *Q* in the *Same Number Choices* in which considering the lifespan's length is trivial or impossible.

Furthermore, in cases like Alexander or the Climber, whoever makes a choice is not sure of its outcome. For example, the Climber knows that she has more probability of dying if he chooses to climb difficult mountains, but she does not know which escalade will kill her: the fatal escalade might be the next, or it might never happen. The choice of climbing difficult mountains has a great variety of possible outcomes, and it is impossible to know which one will become reality. Thus, finding *Theory X* might be more useful for solving cases like *Alexander's choice* if *Theory X* is able to treat uncertainties, more specifically the uncertainty about the difference of quality and duration of a life.

It might be argued that, actually, the line between *Same Number* and *Different Number Choices* should disappear at some extent. It might be stated in fact that, since in an alternative the Climber lives longer, there will come a time when her choice will have made a difference in the number of existing people: if she chooses *CLife 1* there will be one person less than the alternative, and there will exist one more person than the alternative if she chooses *CLife 2*. Choices like *the Climber's choice* can be considered somehow between *Some Number Choices* and *Different Number Choices.* Since the greatest majority of *Same Number Choices* would be like *the Climber's choice*, if we knew all the facts, it might be stated that the greatest majority of *Same Number Choices* are actually *Different Number Choices*. Still, I

prefer to maintain Parfit's distinction. Interpreting the greatest majority of *Same Number Choices* as akin to *Different Number Choices* seems to me perfectly equivalent to classifying them in a new group which has features alike to *Different Number Choices*. Classifying them directly in those latter seems to me a too strong claim, since it is not clear to what extent the length of a life can be made equal to the number of existing people in the history of a population.

Before proceeding, it might help to recap quickly why the majority of *Same Number Choices* and all *Different Number Choices* are extremely similar. Both in *Same Number Choices* like *the Climber's Choice* (that are the majority of *Same Number Choices*) and in *Different Number Choices* it is unclear whether adding more life (added extending someone's lifetime or causing to exist further members of a population) is worth the sacrifice of lowering the average welfare. In both kind of choices *Q* is unfit, and *Theory X* seems necessary. It is now time to show what features such a *Theory X* must have. I will do so by pointing out (1) what features are required for a theory fit for Same Number Choices, then remembering (2) what features are required for a theory fit for *Different Number Choices* (that is to say, a good theory for Population Ethics) and finally by merging (1) and (2).

For solving cases like *Alexander's Choice*, which are *Same Number Choices* and might be the most frequent kind of moral choices,[378] we need an adequate theory of choice value where an individual's [a] amount of lifetime moments, [b] welfare and [c] identity may vary. Furthermore, this theory should [d] be able to treat uncertainties regarding [a] and [b].

I use the term "(an individual's) amount of lifetime moments" instead of "lifespan" because the former has more flexible features than the latter. In fact, the term "lifespan" refers to all duration of a life, which is nothing else than the set of successive moments in an individual's life that goes from her birth to her death. But I cannot use the term "lifespan" in sentence like "I add a lifespan of positive welfare in someone's life", whereas I can use the expression "I add a certain amount of lifetime moments with positive welfare in someone's life".

I already defined welfare, and explained why identity may vary.

[378] I remember that *Alexander's Choice*, like *The Climber's choice*, is a moral choice since it affects the existence of a future being.

By restating the definition of population Ethics, the similarity between the requirements for a theory that helps in *Same Number Choices* and the requirements for a theory that helps in *Different Number Choices* should become clear. Population Ethics is the field of moral reasoning that seeks "an adequate theory of population value where [a] the number of people, [b] their welfare, and [c] their identities may vary". I now unify the requirements for solving cases like *Alexander's Choice* and the objectives of Population Ethics.

Populations must be conceived, for practical purpose, as something that is the result of a choice. It is true that a Population Axiology might be used for describing populations, but a Population Axiology is a tool that, for the aims of Population Ethics, has its main utility in offering the criteria according to which the best population has to be *chosen*. Aiming at choosing the morally best population means aiming at ranking every choice that concerns populations as to its morality. It might be stated that a Population Axiology is the basis for an axiology of choices between population, and this latter axiology is the main interest of Population Ethics. Therefore, Population Ethics can be described as a theory of choice value as well as a theory of population value.

In actual choices, anyone that performs *Different Number Choices* cannot foresee exactly the consequences of the choice. A choice might have different possible outcomes, especially if it is a choice whose consequences extends greatly in time.[379] It is important to be able to treat uncertainties in *Different Number Choices* as it is in *Same Number Choices*. Therefore *Theory X* must be able to cope with uncertainties.

The term "welfare" is fit for both *Same Number* and *Different Number Choices*. The welfare enjoyed by a population depends certainly on the welfares of the persons living in it.

Someone might try to equalize "number of future people" and "(an individual's) amount of lifetime moments" trough Parfit's concept of Personal Identity, but it would probably be impossible. If Parfit is right, in a life there is a succession of different selves, but this does not help, since they *cannot* be numbered

[379] This has been particularly stressed in (Broome, 2006, p. 78-86 particularly)

and treated as sets composing a person's lifetime.[380] The idea of stating that populations are sets of lives, and lives are set of successive moments in an individual's life, seems way more appealing. If this idea is right, we can conceive the number of people's lives in *Different Number Choices* as the width of a population, and the width of a population as the sum of the set of successive moments of all individual's life. If we believe in Parfit's view of Personal Identity there is no difficulty in accepting this view. It must nevertheless be admitted that Parfit's view is counter-intuitive at first sight and that, if this counter-intuitive view is refused, the conception of a population as the simple sum of the set of successive moments of the individual's life might be controversial. This aspect might need to be more carefully considered, but those considerations will not be done it in this work.[381] Since this is so, "number of people" and "time lived (in someone's life)" will be left separated.

I can now say that the best account for Beneficence, or *Theory X*, ought to be a theory of choice value where [a] the number of people, [b] their welfare, [c] their amount of lifetime moments and [d] their identities may vary. Furthermore, it should [e] be able to treat uncertainties regarding [a], [b] and [c].

V is reliable only for *Same People Choices*, but very few of those kinds of choice are morally relevant. *Q* is reliable for some *Same People Choices*, and if we knew how our choices affect our lifespan it would be reliable for very few choices. Also, *V* is reliable only because *Q* coincides with it in *Same People Choices*, and *Q* is reliable only because *Theory X* coincides with *Q* in some *Same Number Choices*. Therefore, the importance of *Theory X* is paramount. *Theory X* is the only reliable principle for *Different Number Choices* and for *Same Number Choices* in which it is important to consider the length of a person's lifespan. In both cases, it would be helpful if such a theory would be able to treat uncertainties.

It must be highlighted that *Theory X*, the theory at which Population Ethics' thinkers aim, must work also on *Same Number Choices*. This is so because the problem raised by *Same Number Choices* is extremely alike to the problem raised by

[380] Remember the *Psychophysical Spectrum*: as it is impossible to understand what small change in features separates me now from Greta Garbo at the age of 30, it is impossible to understand exactly *when* it appeared a self successive to my 6-year-old self, even if there is no doubt that I am a different self than my 6-years-old version, since I am not psychologically connected with him. Asking when a self becomes a later self is an empty question. Since this is so, the selves in a life of a person cannot be numbered.

[381] A discussion of this kind can be found in (Broome, 2006, p. 104-116)

Different Number Choices. This means that, if a proposal of *Theory X* is unfit for *Same Number Choices*, it must be discharged. On the opposite, it is possible to state that if a proposal for *Theory X* can be considered effective for cases like *The Climber's choice*, which has features extremely alike to the features of any *Different Number Choice*, it should be likely to be convincing also for *Different Number Choices*.

In Population Ethics three alternative conclusions have been reached so far. Some thinkers believe that *The Repugnant Conclusion* has to be rejected, and that *Theory X* is a different principle than the *Impersonal Total Principle*. Some other thinkers believe that *The Repugnant Conclusion* can be accepted, and therefore the *Impersonal Total Principle* is *Theory X*. Others believe that a satisfactory Population Axiology cannot be found.

It is worth to spend a few words on this latter position. In particular, we will refer to the already mentioned Gustaf Arrhenius, one of the most influential philosophers in Population Ethics and holder of the latter view. He found strong arguments against many important proposals of Population Axiology and of *Theory X*, if not against *all* important proposals.[382] Furthermore, he listed some adequacy conditions that a Population Axiology (and consequently also a *Theory X*) should respect. Arrhenius believes uncontroversial those adequacy conditions.[383] He has good reasons for believing it: many theorists accept all those conditions, and all theorists accept the majority of them.[384] Arrhenius showed, through theorems demonstrated like Arrow's theorem,[385] that those uncontroversial adequacy conditions are incompatible.[386] Since this is so, a Population Axiology that satisfies them all cannot be found. If such an Axiology cannot be found, we cannot find a

[382] He criticized *Total* and *Average Principle* in (Arrhenius, 2000, p. 37-57) and he objected a great number of the other most influential theories in (Arrhenius, 2000, p. 58-150).
[383] He stated that they are "to the best of [his] knowledge, the logically weakest and intuitively most compelling conditions" (Arrhenius, 2016) for a satisfactory *Theory X*.
[384] One of those adequacy conditions, namely the *Quality Condition*, will be objected in 9.2 .
[385] Kenneth Joseph Arrow (1921-2017), American economist winner of the Nobel Prize for Economy in 1971, demonstrated in his *Social Choice and Individual Values* (1951) that, if who votes can choose between more than three options, there cannot be a ranked voting electoral system that can both mirror the ranked preferences of individuals and respect some very compelling principles of adequacy, such as the avoidance of dictatorship.
[386] Arrhenius writes his theorems in different books. Six of them can be found in (Arrhenius, 2000, p. 151-198), others can be found in (Arrhenius, 2003), (Arrhenius, 2009) and (Arrhenius, 2011) (this latter is actually a revision of (Arrhenius, 2009)).

satisfactory *Theory X* either. If such a theory cannot be found, we should abandon some of Arrhenius' adequacy conditions or become moral skeptics, accepting that our moral beliefs are not epistemologically justified.[387]

It might be stated that, despite the unappealing epistemological problem, a great part of morality can still rely on *V* and Q, thus we are not obliged to become moral skeptics. Moral skepticism would be avoidable due to the abstractness of the problem: since the problem is, in the greatest majority of the cases, merely theoretical, it seems that we have no reason to become skeptics in practice. Unfortunately, Arrhenius is right in stating that revising some adequacy conditions for Population Axiology and moral skepticism[388] are the only two alternatives. It has been tried to demonstrate in this section that *V* covers only few relevant choices (the ones in which the *only* person that is benefited is close to death), and *Q* is valid only in choices in which (1) the difference in lifespan of the outcomes is trivial and (2) the difference in lifespan of the outcomes is impossible to estimate. It is reasonable to think that choices of the kind (2) are much more frequent than choices of the kind (1). Choices of the kind (2) can be treated by *Q* only because we lack of an important information. If no *Theory X* can be found, we should be grateful for our ignorance, since we can rely on *Q* in cases (2), which are many. But this would mean that a remarkable number of problems in normative ethics can be solved only if we ignore some relevant fact. A theory that is effective only when relevant facts are unknown is not a reliable theory, and it would be rational to distrust it, and become skeptical about it.

If thinkers such as Arrhenius are right and no satisfactory *Theory X* can be found, morality seems to be under great threat. We must admit that the majority of

[387] (Arrhenius, 2000, p. 200). There is actually a third option, that is explaining away the results of population Ethics in morality. This possibility will not be explored here, even if it is worth noticing that Arrhenius hypothesized that "a closer investigation of [the just stated] option might bring to light an as yet unforeseen solution to the problems discussed in this essay. Perhaps we should not take the paradoxes of future generations as a challenge to the existence of a satisfactory moral theory, but as a challenge to some of our beliefs about moral justification and about the purpose and scope of moral theory." (Arrhenius, 2000, p. 202). An example of research in this sense is to try to reject the assumption according to which the relation "a population X is better than a population Y" is transitive. The fact that "a population X is better than a population Y" is transitive is usually called "principle of Transitivity". If this principle can be rejected, then if population X is better than population Y and population Y is better than population Z, it can be admitted that population Z is better than population X. This is supported by arguments, for example, in (Temkin, 2012). Parfit's last article on Population Ethics, which is (Parfit, 2016), supported this opinion too.

[388] And the rejection of the principle of Transitivity. See note 387.

morally relevant questions cannot be answered adequately, or reject some apparently plausible condition instead. Since some thinkers refuse the adequacy conditions that lead to the rejection of *The Repugnant Conclusion*, it is worth to explore the possibility for refusing those conditions by examining if *The Repugnant Conclusion* is acceptable in *Same Number Choices*. In fact, if the similarity of cases such as *The Climber's Choice* and *Different Number Choices* has to be admitted, and if *The Repugnant Conclusion* can be admitted in those former cases, it is plausible to claim that it has to be admitted also in those latter cases. This possibility will be explored in chapter 9.

Chapter 9: Population Ethics applied to an individual's life

In all cases of *Same Number Choices* we can know only approximately, at best, how the choice will affect the amounts of lifetime moments and the quality of life of future selves. Some considerations about the possibility of a *Theory X* that can treat uncertainties will be made in chapter 10. Now, for simplicity, let us assume that those values can be known precisely. This will let us understand better if a principle of Population Ethics is fit or not for solving *Same Number Choices*.

It is first the case to answer a possible objection. I stated that a theory that solves *Different Number Problems*, a *Theory X*, must be fit also for *Same Number Choices*. I am now proceeding to test if some adequacy condition for *Theory X*, for example the avoidance of *the Repugnant Conclusion*, is fit for *Same Number Choices*. Those test will involve, for example, establishing if we find the analogous of *the Repugnant Conclusion* in *Same Number Choices* still repugnant, or if we find that an analogous of the *Non-Sadism Condition* is fit or not for a theory that helps in *Same Number Choices*. It might be objected that those kinds of test have no value for *Different Number Choices*. In fact, it might be possible, for example, that *the Repugnant Conclusion* is acceptable in *Same Number Choices* but not in *Different Number Choices*, and therefore we would need a *Theory X* that allows such a conclusion in *Same Number Choices* and avoids it in *Different Number Choices*. Since this is so, evaluating a *Theory X* merely through tests in *Same Number Choices* might be trivial when considering *Different Number Choices*. An adequacy condition fit (or unfit) for how *Theory X* behaves in *Same Number Choices* might be unfit (or

fit) in *Different Number Choices*, and thus this whole chapter would have no value when asking what are the adequacy conditions for a *Theory X* that helps *Different Number Choices*. Consequently, this whole chapter would have no value when deciding whether we should reject some of Arrhenius' adequacy conditions for a Population Axiology or we should become moral skeptics.

It might be true that *Theory X* may allow different things in *Different Number* and *Same Number Choices*. If there is such a discrepancy, then *Same Number* and *Different Number Choices* are not as similar as so far stated. Since this is so, when in the following arguments there might be discrepancies between how *Theory X* should behave in the two kinds of choices, I will suggest how my evaluations in *Same Number Choices* are very likely to have correspondences in *Different Number Choices*.[389] Still, it might be argued that the correspondences I suggest are somehow biased: it can be claimed that, when making a *Same Number Choice*, we can identify ourselves with the person that will be affected by the consequences of the choice, and this may lead us to make some kind of evaluation that has our conception of ourselves as premise. This conception of ourselves cannot be hold also by a population, since a population cannot have a conception of itself, and therefore an evaluation of a theory in *Same Number Choices* would be unfit for evaluating *Different Number Choices*. In other words, since no one can identify herself with a population, it might be argued that we cannot evaluate the fitness of a theory that helps in *Different Number Choices* basing on its application on single individuals.

I agree with the statement according to which we cannot identify ourselves with a population. But I believe that someone cannot even identify herself with *her whole lifetime*. I cannot identify myself with what I am not psychologically connected with. If I think about my lifetime, I know I am not able to evaluate it reliably. In fact, in any moment of my life, when I think about my lifetime I give more weight to the moments with which I am more connected. It is inevitable to do so. For example, if I do not remember a moment of my lifetime, I cannot give to this moment its proper weight, even if that event may have actually a great weight. This is plausible: according to some psychologists, some of our memories are removed,

[389] It will happen only once, in 9.2, when evaluating *the Repugnant Conclusion*. *The Repugnant Conclusion* will not seem obviously repugnant in *Same Number Choices*, while, according to someone, it is repugnant in *Different Number Choices*.

but they still influence us.

Even if I remember a moment, I may give this moment an evaluation very different accordingly to when I remember it. I may be happy to have had some experiences while I am young, since I enjoyed them, I may disapprove the same experiences when I will be older, since I will find them irresponsible. In at least one of the two moments I am giving the wrong weight to those experience. It is likely that there is no time in our lifetime in which we can give the proper weight to every moment we lived. It seems reasonable to claim that we unavoidably give more weight to the moment we are more connected with.[390]

If this is true, we should not let any conception of ourselves influence us when evaluating a *Same Number Choice*. Certainly, when we try to predict what will be best for our future selves we *need* to know ourselves and have some conception of who we are, because ourselves at the present time are the best people for guessing the features of our future selves: in this sense, our conception of ourselves can influence a *Same Number Choice* that concerns ourselves. But when we evaluate theoretically a possible principle for *Same Number Choice* we do not rely on conception of ourselves. Therefore it is unlikely that, when I suggest correspondences between *Same Number Choices* and *Different Number Choices*, they will be somehow biased by considerations on ourselves.

Let us start by seeing if the two proposals of *Theory X* so far stated in this work, that are the *Total* and the *Average Principle,* seem fit in *Same Number Choices*. The *Average Principle* will be discharged in 9.1 quickly while, as I anticipated, in 9.2 the *Total Principle* will find some support.

9.1. *The Non-Sadism Condition* in *Same Number Choices*

The strongest objection against the *Impersonal Average Principle*, according to which the best outcome of a moral choice is the outcome in which lives go on average best, is probably Arrhenius' observation that this principle violates, in *Different Number Choices,*

[390] If I am right, it is incorrect to state sentences like "my life was worth living so far" or "my life was not worth living so far". Such kind of sentences can only be intended as "I evaluate the moments of my life with which I am connected as worth living/not worth living. The strength of my evaluation for each moment is directly proportional to my connectedness with that moment".

> *The Non-Sadism Condition*: an addition of any number of people with positive welfare is at least as good as an addition of any number of people with negative welfare, other things being equal.[391]

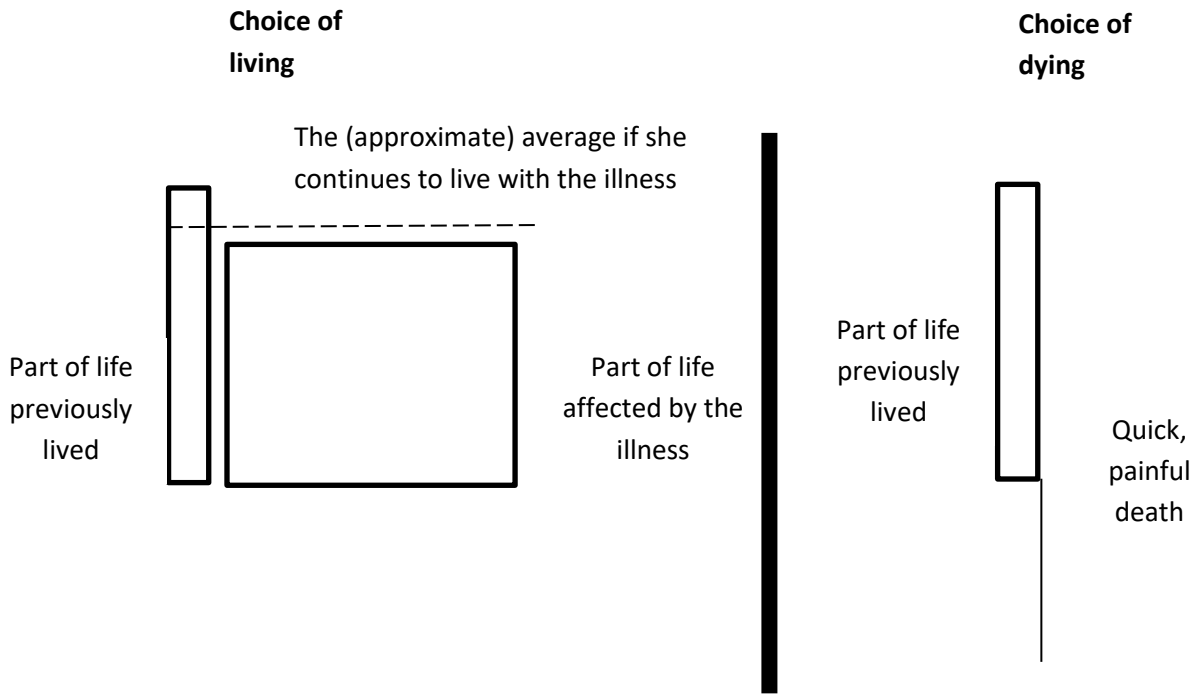
We expect to find that the *Impersonal Average Principle* makes a similar violation in *Same Number Choices*. If we cannot find such a violation we might suspect that *Different Number Choices* and *Same Number Choices* are not so alike as previously stated. The two kinds of choices can be treated alike if the same principle is similarly unfit for both choices; if we find that a principle seems acceptable in one kind of choice but inacceptable in the other kind it would be unclear how far *Same Number* and *Different Number Choices* can be treated similarly.

Imagine a *Same Number Choice*. Let us call it

> *The Illness*. A young person has so far lived a life that is very worth living. She now discovers that she has an incurable illness that will predictably affect her future life for the worse, but not too much for the worse. Her life would probably be very long and will be still very worth living, but less worth living than what she has lived so far. She can chose to live with this illness or, instead, have a quick, but painful death. Whatever her decision will be, she will be supported by everyone loves her and is loved by her, and her death will not be grieved in a different way whatever her decision, since everyone trusts that the mentioned decision will be guided by the best principle for *Same Number Choices*.

The squares below show how much less her life would be worth living in her future ill life and how painful would be her death. I make the same assumptions I did for the case of the Climber, except the assumption that only climbing makes life worth living: the width of the squares is the length of her life, the height of the squares represents her welfare.

[391] (Arrhenius, 2000, p. 64, 203).



In tracing those blocks it has been assumed that a person can be better off or worse off if she makes a choice rather than another. It has been assumed that a life is considered having welfare 0 if and only if it has the following feature: if someone could choose between living or not living such a life, she would be indifferent. It has also been assumed that alternative welfares can be measured and compared one another. For simplicity, it is assumed that the lifetime so far lived by this person has the same welfare at any time, and the future time this person can live will have a value of welfare that does not change at any time. It is also assumed the correctness of what written in 8.2 about future identities, which is that any choice that affects the future identity of an existing person is *not* a *Same People Choice* (unless it is a morally trivial choice or it concerns a dying person) and therefore those choices cannot be made relying on *V*.

If she relies on the *Average Principle*, she would choose to end her life. In fact, if she chooses to end her life, the average welfare she would have enjoyed would be roughly equal to the welfare of the part of her life previously lived. I write "roughly" because her painful death affects for the worse the average welfare of her life, but since her death is not a long event it will not lower it significantly.

If instead she decides to live, the average welfare would be at the level of the dashed line, which is lower than the average welfare of the part of life previously lived, and therefore is lower than the alternative in which she kills herself.

Many will find that choosing dying rather than living a good life, even if this life is a bit worse than what lived so far, is not a good choice. Someone might state that her tragic choice is a too dramatic reaction for the news of her illness.

It is therefore possible to state that, for *Same Number Choices*, the *Average Principle* is unfit, since it violates an adequacy condition similar to the *Non-Sadism Condition*, which can be defined as

> *The Non-Dramatic Condition*: a prolongation of lifespan in which the life is worth living is at least as good as a prolongation of lifespan in which the life is not worth living.[392]

We found an analogy in the objections to the *Average Principle* raised in *Same* and *Different number choices*.

9.2. *The Quality Condition* in *Same Number Choices*

How about the *Total Principle?* Is there a case akin to *the Repugnant Conclusion* in *Same Number Choices*? According to Derek Parfit,[393] such a case can be imagined. The case is merely hypothetical, as the Repugnant Conclusion is. He writes:

> Consider […] the analogue, within one life, of the Repugnant Conclusion. Suppose that I can choose between two futures. I could live for another 100 years, all of an extremely high quality. Call this the Century of Ecstasy. I could instead live for ever, with a life that would always be barely worth living. Though there would be nothing bad in this life, the only good things would be muzak [which is awful music] and potatoes [which Parfit uses as an example of a not really fine food]. Call this the Drab Eternity. I believe that, of these two, the Century of Ecstasy would give me a better future. And this is the future that I would prefer. Many people would have the same belief, and preference. On one view about what makes our lives go best, we would be making a mistake. On this view, though the Century of Ecstasy would have great value for me, this value would be finite, or have an upper limit. In contrast, since each day in the Drab Eternity would have the same small value for me, there would be no

[392] In *the Illness* her death is considered as a prolongation of her life; more specifically, the life is prolonged by the amount of time she uses for killing herself.
[393] And McMahan, because what follows is partly based on an unpublished paper, as stated in (Parfit, 2004, p. 17, note 14).

> limit to the total value for me of this second life. This value must, in the end, be greater than the limited value of the Century of Ecstasy. I reject this view. I claim that, though each day of the Drab Eternity would be worth living, the Century of Ecstasy would give me a better life. This is like Mill's claim about the 'difference in quality' between human and pig-like pleasures.[394] It is often said that Mill's 'higher pleasures' are merely greater pleasures: pleasures with more value. As Sidgwick wrote, 'all qualitative comparison of pleasures must really resolve itself in quantitative [comparison]'.[395] This would be so if the value of all pleasures lay on the same scale. But this is what I have just denied. The Century of Ecstasy would be better for me in an essentially qualitative way. Though each day of the Drab Eternity would have some value for me, no amount of this value could be as good for me as the Century of Ecstasy.[396]

Arrhenius tried to make more explicit what is the feature that Parfit finds repugnant in *the Repugnant Conclusion* in *Different Number Choices*. According to Arrhenius, the intuition behind Parfit's *Repugnant Conclusion* is that the *Total Principle* violates what Arrhenius calls

> *The Quality Condition*: There is at least one perfectly equal population with very high welfare which is at least as good as any population with very low positive welfare, other things being equal.[397]

The Quality Condition, as *The Non-Sadism Condition,* is part of a list of adequacy conditions that, in Arrhenius' opinion, *Theory X* should not disregard. This adequacy condition, as every Arrhenius' adequacy condition, seems very compelling, since it points out how counter-intuitive seems to be the fact that any population with very high welfare is worse than some population with very low welfare. Arrhenius adds the following example in order to support this adequacy condition:

> Let's assume that […] the current world population consists of people with very low positive welfare. Which of the following two futures would be the best? In the first scenario we have a massive expansion of the population size but all the people still have very low positive

[394] (Mill, 1861, p. 188 and generally part II)
[395] (Sidgwick, 1874, p. 94)
[396] (Parfit, 2004, p. 17-18)
[397] (Arrhenius, 2000, p. 41). This condition is the basis for another condition: *The Quality Addition Principle* (Arrhenius, 2000, p. 55, 56 in its weak version). Any successful objection to *The Quality Condition* undermines *The Quality Addition Principle* too.

> welfare. In the second scenario, the population size remains the same but we have a major increase in people's welfare such that everybody enjoys very high welfare. The answer seems obvious.[398]

Arrhenius assumes that the currently existing population enjoys a low welfare because a frequent argument prompted by supporters of *the Repugnant Conclusion* is stated roughly as follows: on reflection, even the currently existing people with the higher welfare (people living in the "lap of luxury", as Ryberg calls it) have a life that is, on average, barely worth living. We do not find repugnant the idea that more people with the higher existing welfare exist, therefore *The Repugnant Conclusion* is not actually repugnant.[399]

Arguments of this kind can be exemplified by the following lines from (Tännsjö, 2004, p. 223):

> The view that I am prepared to defend is somewhat pessimistic but still, I am afraid, realistic. My impression is that if only our basic needs are satisfied, then most of us are capable of living lives that, on balance, are worth experiencing. However, no matter how "lucky" we are, how many "gadgets" we happen to possess, we rarely reach beyond this level. If sometimes we do, this has little to do with material affluence; rather, bliss, when it does occur, seems to be ephemeral result of such things as requited love, successful creative attempts and, of course, the proper administration of drugs.

This means that, when Arrhenius writes "Let's assume that the current world population consists of people with very low positive welfare", he means "Let's assume that the current world population *that enjoys the highest welfare* consists of people with very low positive welfare".

---

[398] (Arrhenius, 2000, p. 46-47)

[399] Arrhenius reports, in (Arrhenius, 2000, p. 45-46), three arguments of this kind, that belong to Tännsjö, Hare and Ryberg. It is worth to point out that it is not necessary for the supporters of *The Repugnant Conclusion* to believe that the currently existing people with the higher welfare have a life that is, on reflection, only barely worth living. It might be held that life is, on average, worth living, as Sidgwick believed (see 2.3), and that, to live in a better world, some people should sacrifice part of their higher level of welfare in order to reduce the number of people living a life not worth living. For example it seems not repugnant the idea that someone chooses, over having an higher number of cars, a greater house and more expensive vacations (all those can be considered components of an high welfare), having an higher number of children, given that all of them will live a life worth living.

Contrary to *The Non-Sadism Condition*, that to my knowledge has been accepted by every theorist, some authors, such as Huemer, Tännsjö, Broome, Ryberg and Ng, rejected *The Quality Condition* and accepted *The Repugnant Conclusion*.

Parfit's choice between Drab Eternity and the Century of Ecstasy is a *Same Number Choice* that has explicit similarity with *Different Number Choices*. It should therefore be treated alike *the Climber's choice*. When analyzing this latter, we asked if the fact that the climber lived longer made a moral difference. Choosing between Drab Eternity and the Century of Ecstasy, Parfit denied a moral difference caused by the fact that he lived longer. Parfit would be right if and only if in *Same Number Choices* there exist an analogous of *the Quality Condition*. The adequacy condition analogous to *The Quality Condition* in *Same Number Choices* might be called

> *The Valuable Moments Condition*. There is at least one perfectly equal set of successive moments with very high welfare in an individual's life which are at least as good as any set of successive moments with very low positive welfare in an individual's life, other things being equal.

I believe that *the Valuable Moments Condition* is, in *Same Number Choices,* reasonably equal to *The Quality Condition* in *Different Number Choices*. Therefore, if *the Valuable Moments Condition* must clearly be rejected for *Same Number Choices*, we would have strong reasons to suspect that *The Quality Condition* must be rejected too. I will not be able to demonstrate that *the Valuable Moments Condition* is obviously wrong, but I will show how there are good reasons for doubting its truth. Since this is so, there are good reasons for doubting also the truth of *The Quality Condition*.

In order to understand if *the Valuable Moments Condition* is a good adequacy condition for a theory that solves *Same Number Choices*, we need to ask what set of successive moments with very high welfare may be as least as good as any set of successive moments with very low positive welfare. It is not precisely clear what it may be, but it might be enough to know that the set of successive moments that satisfies *the Valuable Moments Condition* must necessarily be composed by a great number of moments, that is to say, a great period of time. In fact no one would think, for example, that a second in which it is enjoyed a very high welfare has to be

preferred to *any* set of successive moments with lower welfare. If someone would sacrifice, for example, the possibility of living three years worthy living in order to live a single second very worth living, she would be labeled as imprudent. Or at least she would be labeled as imprudent if it were certain that she would have enjoyed more welfare during those three years than in the single second of great welfare. No one would try to persuade a loved one to make this kind of sacrifice.[400]

Remember that accepting Parfit's theory of Personal Identity does not let us disregard Prudence. Sidgwick's *Principle of Prudence* prescribed that the moral agent should have impartial concern for all parts of his or her conscious life. A Consequentialist that accepts the unimportance of identity, as suggested by Parfit's theory, would reject the Principle of Prudence. Instead, such a Consequentialist would believe that benefiting a present self at a future self's expense is immoral, because the good of any one individual is of no more importance than the good of any other, as Sidgwick's *Principle of Benevolence*[401] prescribes. Therefore, the sacrifice of years worthy living for a second very worth living is not approved even if Sidgwick's *Principle of Prudence* has been rejected, since this sacrifice is forbidden by the *Principle of Benevolence*.

Parfit knows that the length of the set of successive moments cannot be short, and in fact in his *Same Number Choice* version of *the Repugnant Conclusion* he supposes a hundred years worthy living as set of moments that satisfy the *Valuable Moments Condition*. The set of successive moments that satisfies *the Valuable Moments Condition* must consist in an *enormous* amount of time, as a hundred years are.

---

[400] It might be objected that someone would: in fact, it seems to happen worse. For example, there are people that encourage their friend to take powerful drugs. Those kind of drugs, in exchange for what seems a moment of great welfare, decrease the duration of life and, in addition, makes some moments of the life not worth living. There are two flaws in this objection. The first is that, even if someone encourages her friends to take this kind of drugs, it is very questionable not to consider imprudent this behavior. The second flaw is that those people are not certain that who takes the drugs will have many years worthy living, they do not know if those years worthy living will be denied precisely because they are taking that drug and, in most cases, they do not even consider this possibility. A proof of this is the following: it is very likely that an overdose of cocaine leads to the enjoyment of much of what people who takes cocaine would consider welfare. But an overdose of cocaine has an high chance to kill, therefore no one suggests to take much more cocaine than what guarantees survival, especially if it is certain that the person assuming cocaine will have years worthy living in front of her. This is a clear clue that it does not seem reasonable to sacrifice all future moments worthy living for a single moment very worth living.

[401] See 3.3 or (Sidgwick, 1874, p.382)

Sacrificing a great amount of time worth living, such as three years, for a single second of very high welfare is imprudent, unless maybe if we believe to experience in that second more than what we would experience in three years of life worth living. The same can be said if we sacrifice those three years for a single minute of very high welfare. If four years worthy living are sacrificed for a day very worth living, we still would probably disapprove this choice. Likewise, we would condemn the choice of having a year of high welfare at the cost of losing twenty years worthy living, if in twenty years it is enjoyed more total welfare than in that single day. And so on: with small amounts of time we never believe that sacrificing a long period of life worth living for a smaller period of life very worth living is prudent, if the sum of welfare is greater in the long period than in the short.

Why, then, with a certain greater amount of time, this would change? What is the amount of time which, if very worth living, does not make imprudent any sacrifice of a longer period in which it is experienced more welfare, even if not in the average? A hundred years, writes Parfit. Why? Why not, for example, 70 years? Why not 40? Why not 10? Why not a year, a month, or a second? Where is the line between imprudence and imprudence, and basing on what principle has it been drawn, if *the Valuable Moments Condition* has to be admitted?

According to those who support *the Valuable Moments Condition*, the line that divides imprudence and prudence exists. It is the line before which a length of time with very high welfare does not satisfy *the Valuable Moments Condition* and after which it does. The length of time that satisfies *the Valuable Moments Condition* is not known, but one of its features is known: this length of time is very great. Unfortunately, we have difficulty in grasping how much a cost in time is great for us with length of times that are very great.

Imagine, for example, that you want to build a chair, because you need a chair, you have just enough wood for building a chair and you know that building it is cheaper than buying it. Suppose that you already know how to do it, and therefore you believe that it can be done in two days. You can reasonably believe that a day of your life is an acceptable cost for a chair, and therefore you make your chair in a day. Imagine now, instead, that you have never built a chair, and therefore you need a course of woodworking. The course will be free, but you will be able to build a chair

only after ten days of constant practice: after that, you will be able to build a chair in two days. Therefore, you will invest 12 days in building a chair. If this is the case, you can reasonably decide that 12 days of work are not worth a chair, and therefore you buy a chair, or pay a woodworker for building it for you.

What I am trying to prove is that we believe that some things are worthy of a day of investment, but not twelve days of investment. This means that we can give an approximate value to time of a few days, and compare it with similar times' value.

Instead, imagine that someone wants to have a family. She knows that a family requires efforts and sacrifices, for example it requires her to finding a more stable job than what she had before, even if less enjoyable. If she knew that she will have to make those efforts and sacrifices for 40 years, she might consider those efforts worth of a family. If she knew instead that those efforts would be to endure for 50 years, given that she knows that that she will have the energy for performing them, she might believe that they are still worth it. The same would be if the efforts are required for 70, or a hundred years.

In other words, *any* effort that is considered worth doing for, let us say, 40 years, it is also probably considered worth doing for indefinitely more time. I can think of no sacrifice that anyone would perform for 40 years, but not for a hundred years, given that she has enough energy for this effort. There seems to be *no* things that are worth of a sacrifice of 40 years of someone's life, but not of a hundred years of life.

It might be objected that my example works only because we believe that a family may be worth of any kind of investment in time. It is true, but let us check other examples of things that may be worth an investment of 40 years. Such things may be ideals, scientific researches and artworks. All three can be considered worthy of any investment in time, given that who invests the time will have enough energy for all this time. It must be admitted that *any investment that seems worth 40 years of sacrifices seems worth any investment in time by an individual.*

This means that the value we give to a great amount of time is so approximate that we cannot compare this value with times even much greater. Since this value is so approximate, we have to admit that we have difficulty in evaluating great amounts of times.

To sum up: the amount of time very worth living that satisfies *the Valuable Moments Condition* is not precisely determined, but it is very high. Furthermore, it is hard to evaluate very high amounts of time.

This means that the amount of time with high welfare that satisfies *the Valuable Moments Condition* is a time that we cannot evaluate correctly. This is probably part of the reason why we cannot point out precisely when it is imprudent to sacrifice a long period of life worth living for a smaller period of life very worth living, if the sum of welfare is greater in the long period than in the short.

But it seems very questionable to claim that there exists a time we cannot evaluate that, if it has a very high welfare, it is more valuable than much greater times (which therefore we cannot evaluate as well) with lower welfares. A time we cannot evaluate is compared with other times which we cannot evaluate, and it is stated that a rule valid for times we can compare is not valid anymore. This rule is the one according to which it is imprudent to sacrifice a long period of life worth living for a smaller period of life very worth living, if the sum of welfare is greater in the long period than in the short.

I do not find uncontroversial a condition built this way. *The Valuable Moments Condition* states that a rule valid for times we can evaluate, even if roughly, is not valid with times that we cannot evaluate at all. *The Valuable Moments Condition* must be demonstrated, and not posed as uncontroversial adequacy condition.

Since *the Valuable Moments Condition* appears in this work for the first time, no one is guilty of having overlooked its demonstration. But this condition is equivalent to Arrhenius' *Quality Condition*. Arrhenius believed it to be uncontroversial, but maybe it was only very appealing (or, maybe, simply more appealing than *the Repugnant Conclusion*). It has been showed how we have difficulty in evaluating enormous amounts of times. Likewise, we cannot conceive enormous populations.

The fact that *the Valuable Moments Condition* is not uncontroversial is not a decisive argument against *the Quality Condition*: there might be a discrepancy between *Same Number* and *Different Number Choices*, that is to say, it is possible

that something is valid in *Some Number Choices* and something else is valid in *Different Number Choices*. But if *the Valuable Moments Condition* is not uncontroversial, we have reasons to suspect that *the Quality Condition* is not uncontroversial, too. In fact, for example, it can be admitted that we should prefer adding two persons whose lives' sum of welfare is higher than the welfare of a single person's life we could add instead, and whose welfare is higher than the average of the two persons' lives.[402] Similarly, we admit that it is preferable adding ten persons whose lives' sum of welfare is higher than the welfare of three person's lives we could add instead, and whose welfare is higher than the average of the ten persons' lives. We admit that it is better adding fifty persons whose life have an higher total welfare than ten persons' lives, and so on. Why should this not be admitted with populations that, being much bigger, are harder to conceive?

The Quality Condition and *The Repugnant Conclusion* are incompatible. One of them must be rejected and the other accepted: the alternative seems to be giving up entirely morality.[403] If we have to give up morality entirely because of a condition, this condition must be beyond every doubt. A condition that that might be built on our difficulty of grasping high numbers is *not* beyond any doubt.

Parfit claimed that "though each day of the Drab Eternity would be worth living, the Century of Ecstasy would give me a better life. This is like Mill's claim about the 'difference in quality' between human and pig-like pleasures". He also claimed and that the value of pleasure does not lay on the same scale.[404] I have no decisive proof that value of pleasure lays on the same scale, even if I find Sidgwick's account, according to which difference in quality of pleasures is merely difference of quantity,[405] at least as reliable as Parfit's. But let us admit that there is difference in quality between human and pig-like pleasures. This would mean that it would be right to sacrifice all of pigs' pleasures for a single moment of the greatest pleasures of a man. Are we sure we believe that we can experience a pleasure so

[402] Not admitting it leads to implausible conclusions, such as "*The Reversed Repugnant Conclusion*: For any population with very high positive welfare, there is a better population consisting of just one person with slightly higher welfare, other things being equal." (Arrhenius, 2000, p. 54), or the fact that it would be always wrong to cause a child to exist if the child's existence lowers the family's average welfare.
[403] Or the Principle of Transitivity. See note 387. It can nevertheless be guessed that giving up the Principle of Transitivity will lead to give up morality, too.
[404] See the beginning of the present section.
[405] See 2.3.

great that it would be worth denying all swines' positive welfare for a single moment of this pleasure? We would be, basically, pigs' Utility Monsters. The claim according to which a moment of the highest human pleasures is worth the sacrifice of any amount of pig pleasures seems at least as plausible as its denial.

Return now to how Arrhenius supported *The Quality Condition*. He stated that, assuming that the current world population consists of people with low positive welfare, we should prefer that the already existing population increases welfare rather than a massive expansion of the population size. An equivalent of Arrhenius' statement in *Same Number Choices* will not be attempted. Instead, let us suggest that Arrhenius, in his argument for supporting *the Quality Condition*, confounds two questions. He states that the answer to the question "what is the best future between a future in which everyone has the same low welfare as today in the luckiest part of this word but there are much more people, and a future in which there is as much people as today but everyone has much more welfare?" is obvious. I believe he is mistaken. What is obvious is rather the answer to the question "where would you prefer to live between the two futures?". Everyone would prefer living in the population with the highest average welfare, but this implies not that the population with the highest average is better than the alternative. This is so because, if I knew that in the future in which everyone has an high welfare I would not exist, I would be *indifferent* regarding this future: in Arrhenius' example it is taken as granted the existence of the one who makes a choice. Once assumed the existence of who chooses, the case Arrhenius' considers is influenced by considerations that have nothing to do with *the Repugnant Conclusion* and Population Ethics.

It might be stated that denying *The Quality Condition* is not enough for showing that *the Repugnant Conclusion* has to be admitted: it has to be shown first why this conclusion is not repugnant. Actually, several authors provided arguments for accepting it. It has for example been pointed out that a life barely worth living is something desirable, and not something to refrain from;[406] it has been noticed that we have difficulty in imagining great numbers, therefore our feelings towards a future with a very large number of people, such as the one of *the Repugnant Conclusion*,

[406] See (Huemer, 2008, p. 11,12), (Tännsjö, 2004, p. 224-226), (Ryberg, 2004, p. 239-242) and others.

are not to be trusted without reflection;[407] it has been remarked that we tend to privilege already existing people (for example discouraging a couple to have a further child, but being happy for the child's existence after the child is born) therefore we would probably be happy of living in *the Repugnant Conclusion* once we live in it, even if it might now seem unappealing.[408] Many other arguments have been added. I leave to the bibliography the task of providing them.

I will not give arguments against the supposed unacceptability of the Repugnant Conclusion. Still, I suggest why the Repugnant Conclusion might actually be appealing. It might be argued that people in the richest part of the world live nowadays lives that are more than barely worth living: those people have many cars, go to holidays, live in large houses and eat three or four times a day. Not everyone is that lucky: there are (1) some people living lives not worth living and (2) people that have not been generated, since their existence might have not been worth living. It might not be unreasonable to hope for a future population in which no one enjoys the high welfare of the richest nowadays living on this planet, but instead everyone has a life worth living, even if barely, and there exist many more people, each with a life worth living. Sacrificing the existence of unnecessary goods, such as many cars possessed by a single owner and holidays, might be a reasonable price to pay for allowing everyone a worth life. This sacrifice may seem even more reasonable if the number of future people with a life worth living is significantly higher than the number of currently existing people.

It is now time to examine the relevance of uncertainties in Population Ethics.

Chapter 10: Expected utility

In order to apply in actual moral practice the concepts of Population Ethics, we need to understand how to treat uncertainties. Some Consequentialist thinkers coped with

[407] Huemer illustrates this point with the following jest: "An astronomer giving a public lecture mentions that the sun will burn out in five billion years. An audience member becomes extremely agitated at the news. The lecturer tries to reassure him: 'No need to worry, it will not happen for another five billion years.' The audience member breathes a sigh of relief, explaining, 'Oh, five billion years. I thought you said five million years!" (Huemer, 2008, p. 9). The fact that we have difficulties is evaluating large numbers is also an important point of (Parfit, 2016) and, somehow, of the present part of this work.

[408] See (Huemer, 2008, p. 16-18)

uncertainties through a function of expected utility,[409] and it will be done here too. Here is an example of this function.

Call choices with capital letters *A, B, C* and so on. Call the possible histories that are outcomes of such choices Ah_1,Ah_2,Ah_3…Ah_n, Bh_1,Bh_2,Bh_3…Bh_n and so on. Call the welfare of each history WAh_1,WAh_2,WAh_3…WAh_n, WBh_1,WBh_2,WBh_3…WBh_n and so on. For any story Yh_x and for each welfare WYh_x of the story Yh_x we can find a number π_x between 0 and 1, that assigns the probability of Yh_x and its WYh_x to be realized. (The sum of all probabilities $\pi_1+ \pi_2+ \pi_3+\ldots+ \pi_n$ equals 1.) Clearly, in actual moral choices, the way any single individual assigns probability will be inevitably imprecise to some extent. The same must be stated for welfare.

The expected utility of a choice *A* is a function so defined

$$\text{Expected utility of } A = \pi_1 WAh_1 + \pi_2 WAh_2 + \pi_3 WAh_3 + \ldots + \pi_n WAh_n$$

If the *Impersonal Total Principle* has to be accepted, the higher the expected utility of a choice, the more the choice has to be preferred. It will be henceforth assumed that the objection to *the Quality Condition* done in 9.2 is enough for reject it, and therefore it will be assumed that the *Impersonal Total Principle* has to be accepted and that *the Repugnant Conclusion* is not in fact repugnant.

Before assuming, in our actual moral choice, that the best choice is the one with the highest expected utility, we have to make another assumption, namely that the function of expected utility is the correct way to cope with uncertainties.

This assumption has to be made more because of lack of a more appealing alternative than because of an actual belief in the reliability of the function itself. In fact, the function of expected utility seems to have flaws when choices have outcomes with an extremely low probability of realization.

For example, imagine a choice K in which there are only two possible outcomes, Kh_1 and Kh_2. The welfare of Kh_1 is $WKh_1 = -10^3$ (remember that a negative welfare indicates the presence of unworthiness of lives, and the more the welfare is negative, the more the lives are not worth living), its probability is

[409] It is used, for example, in (Parfit, 1984, p. 8) and (Broome, 2006)

π_1=0,99.

The welfare of Kh_2 is WKh_2= 10^{9999}, its probability is π_2=0,01. Consider a welfare equal to 10^{9999} to be an enormous welfare. If the expected utility has to be trusted, this choice K would be extremely appealing, since the high probability of a Kh_1, an history with great suffering, is outweighed by the extremely high welfare of Kh_2 despite its derisory probability.

The similarity of this case with *the Repugnant Conclusion* is striking. We can, in fact, formulate what could be called

> *The Perplexing Conclusion*. For any possible choice *X* with high probability of an outcome with high positive welfare, there must be some imaginable choice *K* with an outcome whose welfare is so great that *K* is to be preferred to *X* even if the greatly positive outcome of *K* is extremely unlikeable.

This implication might be an objection to the method of evaluation of rational preference trough expected utility.

It has been claimed that *Same Number* and *Different Number Choices* raised alike difficulties, and therefore they had to be treated alike to some extent. It might be claimed that, since principles for treating uncertainties and *Different Number Choices* raises similar problems, they are to be treated alike too.

It is worth to be cautious: morality and the study rational decisions under uncertainty are completely different fields. The latter can certainly help the former, but to what extent the rules of those fields are alike is not matter of this work.

Despite the difficulties, to my knowledge a better tool than expected utility for rational decision under condition of uncertainty has not been found, while it might be suggested that, in practical choices, extremely unlikeable outcomes could be ignored.[410] Expected utility will be therefore used in this work. What stated in 10.1 will be valid only if expected utility has to be trusted as a function for

[410] The choice of ignoring small probabilities of great gains is easier to accept than the choice of ignoring small probabilities of great damages. For example, it is not clear at all if it might be imprudent, when deciding to build a nuclear plant, not to consider the possibility of a nuclear disaster caused by accident. In my following example, I will leave unanswered the question whether this possibility has to be considered or not.

evaluating rationally preferences under uncertainties. What stated in 10.2 will be valid regardless.[411]

10.1. Expected utility applied

Many supporters of *the Repugnant Conclusion* wrote that we find it repugnant because we cannot estimate correctly when a life is worth living, or because we are not capable of imagining large numbers. They might be right, but there is another fact that makes *the Repugnant Conclusion* unappealing.

When a policymaker performs a choice in real life, she is never exactly sure of what the consequences of her action might be. An unforeseen accident may occur, or the consequences might have been wrongly calculated. More particularly, choices regarding future populations extend so much in time and affect lives so much that they seem to suggest the highest caution.

In such thorny matters, a life that in a policymaker's mind is barely worth living seems to have a high risk to fall below the level where life is not worth living anymore. A policymaker would not choose the existence of many lives barely worth living unless she has strong motives for believing that they will *remain* worth living. The more difficult a life is, the more risk it has of becoming not worth living. Thus it might seem absurd to say that the best possible outcome is the one where there is a huge amount of people whose initial condition is a life barely worth living, because it usually implies a really high risk that the great majority of this people will lower its level of life during time.

This means that a policymaker would choose a solution in which there are many people whose life is barely worth living only if (1) it has not better alternatives, such as a population with the same number of lives but an higher welfare, and if (2) is reasonably sure that the majority of people has not a great risk having lives not worth living. In fact, if such a risk is not reasonably discharged, the lives would be not barely worth living, but will have a negative welfare, even if slightly negative. This would decrease the total welfare, and it is likely that it will be decreased to the point that this total is dramatically negative, since the number of suffering people is

[411] There are other possible difficulties, not discussed in this work, of the application of expected utility in ethics. See (Briggs, 2014).

dramatically high. This might be another reason why the Repugnant Conclusion seems an absurd solution.

It must therefore be admitted that the number of people, their identity and the value of welfare are not the only variables on which the choice of the best population must be evaluated. A choice might have different outcomes, and all must be considered when choosing.

It has therefore to be admitted that a more realistic account of a choice of population ethics in real life will not be between two alternative populations, but more between alternative sets of possible populations.

For example, imagine a policymaker that has to decide if it is the case to build a nuclear power plant in a certain area. The possible choices might be: *A*, which is building the nuclear power plant, and *B*, which is *not* building it. Building a nuclear plant usually attracts workers, therefore choice *A* involves more people than *B*. Let us simplify and believe that *B* has only an outcome, with probability 1, in which things goes exactly as they are always been, and therefore the welfare of people influenced by the policymaker's choice will remain as they are at the moment of the choice. On the opposite *A* might have, for example, four outcomes Ah_1, Ah_2, Ah_3 and Ah_4.

In outcome Ah_1 the nuclear power plant works extremely well for a long time, generating a high welfare AWh_1. The welfare is high also because the nuclear plant has not great side effect, since a disposal for nuclear waste is found before the power plant causes much damages. Outcome Ah_1 has a certain probability π_1.

In outcome Ah_2 the nuclear plant works similarly to how it works in Ah_1, except that the disposal for nuclear waste is found significantly later than in Ah_1 and therefore there will be a welfare WAh_2 lower than WAh_1. The probability π_2 of Ah_2 is likely to be different from the probability of π_1 of Ah_1, but the policymaker might not know precisely the difference between π_1 and π_2. Therefore she would probably approximate and consider π_1 and π_2 even.

In outcome Ah_3 the disposal is not found. The welfare WAh_3 will therefore be lower than the welfare WAh_1 and WAh_2. Perhaps it will be negative, because it is possible that, if the nuclear waste is not disposed, some lives might become not worth living. Its probability π_3 might be lower than probability π_1 and π_2 since the

problem of nuclear waste will be a pressing problem, and it is reasonable to hope that a great number of researches and discoveries will be done. There is still a probability π_3 that those researches will not be enough to find a solution for nuclear waste.

Finally, in outcome Ah_4 the nuclear plant has a dangerous breakdown due to a human error or a natural disaster and kills many people, injures many other people and makes the whole area uninhabitable. The welfare WAh_4 is extremely negative, but the probability π_4 is also extremely low. Since it is extremely low, the policymaker may choose to ignore such an outcome.[412]

Given all that, the policymaker will have to decide whether the benefits of $Ah_1$and Ah_2 are worth the risk of an outcome Ah_3, or the risk of outcomes Ah_3 and Ah_4 if the policymaker chooses to consider the latter. If the theory of expected utility is the best theory for treating uncertainty, the policymaker will choose between A and B by evaluating if the welfare generated by the choice *B* is higher or lower than the expected utility of A, which has to be calculated as follows:

$$\text{Expected utility of } A = \pi_1 WAh_1 + \pi_2 WAh_2 + \pi_3 WAh_3 (+ \pi_4 WAh_4)$$[413]

It is unlikeable that a choice whose consequence extends over the years can certainly have as an outcome only a possible population. A choice will have more possible different outcomes, each with its different probability of realization and different welfare. A Population Axiology may certainly help to understand what is the best set of possible population to choose, but a Population Axiology should be helped with a criterion for treating the different possible outcomes. In this work it is guessed that the criterion for Population Axiology must be the *Impersonal Total Principle* and the criterion for treating the different possible outcomes is the maximization of expected utility. In this way, a Population Axiology is functional for an axiology of moral choices. An axiology of moral choices is maybe the main aim of all normative ethics: normative ethics is in fact prescriptive, meaning that it *prescribes* how an agent should act. Those prescriptions can be done only by stating which choice is the best among the possible choices; in other words, through an axiology of choices.

[412] See note 410.
[413] I put brackets because the policymaker might choose to ignore h_4.

Return briefly to outcome Ah_3 of the choice A. In Ah_3 the nuclear power plant works well, but the problem of nuclear waste is not solved. When nuclear waste will have become a pressing problem all over the planet, there will exist some lives with very low welfare, since the radiations generated by the nuclear waste will affect people for the worse. Those people might not be killed but, on average, the lives of people affected by radiation might be barely worth living, or might be slightly not worth living. A policymaker *cannot* know if lives will be worthy living or not, also because the difference seems blurry. Since this is so, we find repugnant a choice which leads to lives that might be barely worth living: the difference between life barely worth living and slightly not worth living is hard to grasp, and it is impossible to determine precisely how a choice will affect someone, making his life be barely worth living or slightly not worth living. Since lives barely worth living are very close to lives not worth living, we find repugnant that someone creates lives barely worthy living, since any miscalculation might create lives not worth living. Therefore *the Repugnant Conclusion* appears repugnant.[414]

Return now to the formula of expected utility. It works in *Same Number Choices* as much as in *Different Number Choices*. For example, imagine someone that has to decide if study in university or become a sport professional. The first choice might lead to an history Ah_1 with some kind of degree, and a consequent future job in a certain field that gives her earnings for the rest of her life; instead, the first choice might also lead to an history Ah_2 that involves a loss of will of studying and consequently in a career to rethink; the second choice might lead to (Ah_3) a brilliant career and more money until it last, and then a future to build with the money made, or to (Ah_4) an injury that may significantly damage the career, sad and uncommon but not impossible fate of some sportsman. If the person that has to do this kind of choice uses the *Impersonal Total Principle* as basis for a Population Axiology and uses expected utility as a tool for treating uncertainty, he will choose the alternative with the higher expected utility.

Return now to *the Climber's choice* and *Alexander's Choice*. In order to describe *the*

[414] Sidgwick wrote that the *total principle* may appear opposed to common sense "because its show of exactness is grotesquely incongruous with our consciousness of the inevitable inexactness of all such calculations in actual practice" (Sidgwick, 1874, p. 416). If my argument about outcome h_3 of the choice A is correct, the same can be stated about the Repugnant Conclusion.

Climber's choice with a function of expected utility we should describe a possible future for each escalade the Climber performs.

In cases such as *Alexander's Choice*, Alexander needs to understand how probable is that, if he chooses to study medicine, he will like the subjects he studies for all years necessary, which is the same as asking how is it likely that he will finish his studies and work as a physician; he needs to understand how likeable is it to get assumed as a physician, etc etc. Each possibility is a possible history $Ah_1,Ah_2,Ah_3…Ah_n$. Instead, the choice of working on the platform might be, for example, extremely interesting for the developments of the proficiency in chemistry or somehow depressing due to the isolation of the working conditions.[415]

Is it realistic that the sportsman, the Climber and Alexander give an exact evaluation of the welfare and the probability of each possible outcome and calculate the value of expected utility of each choice? No, precision is impossible. The criterion of expected utility is the principle that each should try to apply in their choices. They should consider carefully the probability and the appeal of each outcome and deciding basing on those consideration; the operation of assigning precise numerical values to welfares and probabilities and calculating numerical outcomes is not necessary. An approximation of the outcome is enough.

10.2. Probability and the Utility Monster

Let us consider:

> *Innumerable Lives or Selfish Paradise*. We know that, if we choose a policy, there will be a population with an enormous number of people whose welfare is low, but still positive. The welfare of this population is matched by the welfare of another population that, with another policy, we may cause to happen. This latter population consists in only one person who, due to technologies unknown today, will be more than perfectly happy every day of his life,

[415] Given that Alexander the Physician and Alexander on the Platform might have different lifespans, it might be asked how, even imprecisely, should Alexander estimate what life will allow him to live longer. I think there is not such a method, if not by searching statistics on the length of lives of physicians and workers on the platform. Those statistic data might not exist, but they can certainly be collected. It might be stated that it seems absurd to choose a career basing on statistics on the length of the life of who does that profession. But it must be noticed that, if the difference is great, the difference in life should be taken into account when performing the choice. Probably, the difference of lifespan between who works in a platform and a medic are not relevant, but, for example, the difference in lifespan between who works in the army and who works as a teacher are *relevantly* different, and this difference should be considered when choosing a career.

> enjoying himself alone the welfare of the other possible generation. Due to the technology, this single individual will not suffer solitude or any damage caused by the absence of human contact.

The two solutions, if the focus must be, as the *Impersonal Total Principle* recommends, only on the total sum of welfare, are equivalent. If the two outcomes have the same welfare and we do not consider probability, they are morally equivalent, and thus there is no clue for understanding which population should be preferred. But if, when comparing the choices, we add a reliable theory of probability to the *Impersonal Total Principle*, we might be able to choose.

The population consisting of a single person is less likely to maintain its welfare at the same height, because the entire welfare weights on only a man. If something bad happens to him, the total welfare will significantly decrease, whereas if something bad happens to a small group of people in the former population the total welfare would not decrease much. In other words: a humanitarian catastrophe is more likely when humanity is smaller. Thus, if we do not have any other clue on how to estimate the risk of the two policies, we might choose the former policy. This consideration strengthens the belief that it cannot be wrong to add people with positive welfare. This belief is stated, in a logically weaker form, as adequacy condition for a Population Axiology by Arrhenius.[416] This kind of conditions are the hardest to accept for those who reject *the Repugnant Conclusion*, because a weaker version of the condition according to which it cannot be wrong to add people with positive welfare, if joined by a principle that seems beyond any doubt,[417] implies *the Repugnant Conclusion*, as showed in (Arrhenius, 2000, p. 52), and is in contradiction with *the Quality Condition*, as demonstrated by the first theorem of (Arrhenius, 2000).[418]

---

[416] It is called "*The Quantity Condition*: For any pair of positive welfare levels A and B, such that B is slightly lower than A, and for any number of lives n, there is a greater number of lives m, such that a population of m people at level B is at least as good as a population of n people at level A, other things being equal." (Arrhenius, 2000, p. 51). Arrhenius demonstrates, in (Arrhenius, 2000, p. 52) that it implies *the Repugnant Conclusion* with an extremely appealing principle, called "*The Egalitarian Dominance Condition*: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal." (Arrhenius, 2000, p. 51).

[417] This principle is *The Egalitarian Dominance Condition*. See note 416 for its formulation.

[418] This argument does not imply that any adequacy condition incompatible with the belief that it cannot be wrong to add people with positive welfare must necessarily be discharged in Population

With a simple change of terms, the fact that a lower number of people has more probability to lose welfare challenges also what Nozick called the Utility Monster. This objection to Utilitarianism has been mentioned in 4.1, and will be restated here:

> Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater gains in utility from any sacrifice of others than these others lose. For, unacceptably, the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility.[419]

The Utility Monster might look in some way repugnant because its welfare looks fragile, unstable and therefore not worth investing, for the same reasons of the population consisting of a single individual considered before. Someone may argue that, if the Monster would be immortal, the solution of feeding it at the price of our own wellbeing would not appear less repugnant. At the answer that immortals, living an infinite amount of time, have an extremely high probability of being permanently damaged, and therefore relying on single individuals is too risky, they would probably change the description of this Monster so that it needs to be "filled" by other's sacrifices, but its welfare level cannot be decreased. This objection can be answered as follows.

Imagine that someone must choose between saving the life of a completely unknown child, child that she has no chances to meet again, and saving the life of a completely unknown very old man, man that she has no chances to meet again. Saving one of them means sacrificing the other: one dies, the other survives. She might choose to save the child because, other things being unknown, the child has a higher probability of enjoy more welfare. This justification is, at least, not clearly absurd.

Axiology. I stated in this work that *Theory* X should rely on a Choice Axiology, that is composed by a Population Axiology and a theory of rational use of probability. It is possible that the belief that it cannot be wrong to add people with positive welfare has to be accepted only when considering probabilities, and not in the Population Axiology. It is possible to state that it is never wrong adding people with positive welfare when we are considering probability, whereas this addition could be wrong in Population Axiology, since in this latter field it might be incompatible with more appealing conditions. Nevertheless, I believe that in 9.2 it has been provided a good account on why *the Quality Principle*, that Arrhenius proved incompatible with the belief that it cannot be wrong to add people with positive welfare, has to be rejected in a Population Axiology.

[419] (Nozick, 1974, p. 41)

Things might be less clear in choices between different numbers of people: imagine, this time, that someone must choose between saving the life of a completely unknown child, child that she has no chances to meet again, or saving the life of five completely unknown very old men, men that she has no chances to meet again. In this case choosing the child over the five men might be more complicated. But imagine that she was reasonably sure that the child would enjoy more welfare than the five old men combined. This might be considered a justification at least not absurd for choosing to save the child rather than the men. If this is true, it is not absurd to feed the Monster.

Before concluding this work, it is worth to sum up the main conclusions of the last three chapters.

Once presented the main features of Population Ethics in 8.1, it has been pointed out in 8.2 that many important choices involving the future of an individual have to be considered *Same Number Choices* rather than *Same People Choices*. In 8.3 it has been noticed how some *Same Number Choices* present difficulties akin to *Different Number Choices*, and therefore some *Same Number Choices* need *Theory X* for their solution as well: thus Population Ethics should provide answers also for this kind of choices. Since Arrhenius' impossibility theorems might imply that we have to give up hope for a principle fitting some *Same Number Choices* and all *Different Number Choices*, we checked if the conditions upon which the theorems are built are really unambiguous. Those tests have been done in *Same Number Choices*. The strongest condition against *the Average Principle* seemed confirmed in *Same Number Choices* (9.1), the strongest condition against *the Total Principle* did not work satisfactorily, even if it has not been denied (9.2). Finally, since the one who makes choices is often in a condition of uncertainty, we explored how expected utility might help in finding *Theory X* (10.1) and we found that reasoning on probability seems to encourage the existence of many people (10.2). As a bonus, we might have found an argument against Nozick's Utility Monster.

This final part of the present work leaves at least two great questions unanswered. First, we need a decisive argument against *the Quality Condition*.

Second, expected utility has some great problems, such as what has been called *the Perplexing Conclusion*.

Finally, Population Ethics is a field that can provide moral answers to some crucial practical matters, such as the diffusion of privatization of primary goods, the assignment of priority in healthcare and climate policies. Those are among the most important matters that humanity has to confront with in our present. If what written in chapter 8 is right, Population Ethics can help also in many crucial choices in the life of an individual.

Therefore, much work must be done. In order to make Population Ethics an useful tool both for choices at individual level and at mankind's level, philosophers need not only to solve the theoretical problems involved, but also to ask the collaboration of those who have the competences for applying the principles of Population Ethics to practical life. Economists have already shown their interest, and the contribution of some of them[420] is truly remarkable. Reaching, for example, the help of climatologists, political thinkers and social theorists seems of great importance. Population Ethics seems to be a tool as powerful as deeply it will collaborate with other fields of study.

Conclusion

This work begun with Nietzsche's aphorism §343 from the third book of *the Gay Science*, entitled "on what Cheerfulness signifies". Let us restate it here:

> At last the horizon appears free to us again, even granted that it is not bright; at last our ships may venture out again, venture out to face any danger; all the daring of the lover of knowledge is permitted again; the sea, *our sea*, lies open again; perhaps there has never been such an 'open sea'.

Nietzsche is stating that the death of God let the philosophers dare to rebuild freely everything related to Him, for example morality. Nietzsche has his own ideas on how

[420] I am referring particularly to Yew-Kwang Ng, that had in (Ng, 1989) an important publication, and Partha Dasgupta, whose interest in Population Ethics begins in 1969, which is fifteen years before *Reasons and Persons* and whose (Dasgupta, 2005) can be considered the synthesis of his thinking on this matter. John Broome's contribution deserves at least a mention, too. Still, he has probably to be considered a philosopher as much as an economist. Among his works, I signal the already quoted (Broome, 2006).

morality should be rebuilt. Parfit does not agree with some of Nietzsche's ideas on Morality, even if Parfit himself showed how him and Nietzsche would have probably agreed at a great extent, if they knew the same facts, and how probably not even Nietzsche believed to a great number of his most controversial claims about morality.[421] Nevertheless, Nietzsche's aphorism perfectly states at least two of Parfit's main beliefs. Parfit's entire work might be summarized with these two beliefs. He was probably aware of their relevance for his own thought, since he points out them frequently. Particularly, Parfit concludes every book he writes pointing out both of them at least in the final two pages. These beliefs are what he calls the *Convergence Claim,* that is the core of all three volumes *On What Matters*, and his belief that the next century is crucial for human history. Let us consider them more closely.

Parfit disagrees with Nietzsche on many points about morality. Still, Parfit agrees in stating that Morality has progressed much, and will do much more, due to its detachment from religion. He writes, in the last two pages of *Reasons and Persons*:

> Non-Religious Ethics has been systematically studied, by many people, only since the 1960s. Compared with the other sciences, Non-Religious Ethics is the youngest and the least advanced.[422]
>
> [in the future centuries] there could clearly be higher achievements in the struggle for a wholly just world-wide community. And there could be higher achievements in all of the Arts and Sciences. But the progress could be greatest in what is now the least advanced of these Arts or Sciences. This, I have claimed, is Non-Religious Ethics. Belief in God, or in many gods, prevented the free development of moral reasoning. Disbelief in God, openly admitted by a majority, is a 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.[423]

Parfit was an atheist, but he did not reject religion in morality simply because he believed religious Morality bad in itself. The importance of Non-Religious Ethics is unclear without reading *On What Matters*. We did not examine this work, but its core

[421] See (Parfit, 2011 b, p. 570-606)
[422] (Parfit, 1984, p. 453)
[423] (Parfit, 1984, p. 454)

is strongly supported by this work. The core of *On What Matters* is demonstrating the truth of the

> *Convergence Claim*: If everyone knew all of the relevant non-normative facts, used the same normative concepts, understood and carefully reflected on the relevant arguments, and was not affected by any distorting influence, we would nearly all have similar normative beliefs.[424]

Non-Religious Ethics is the branch of knowledge required in order to reach agreement on Ethics. Ethics needs to be Non-Religious because everyone, of every faith and faithless, has to agree upon Ethics as agrees upon Mathematics. An universally shared Ethics should rely on mere reasoning.

The *Convergence Claim* is more the core of *On What Matters* rather than *Reasons and Persons*: in the former Parfit tries to demonstrate how there are some universally shared moral principles and how Contractarianism, Utilitarianism and Kant's theory can be unified. *On What Matters* is not concern of the present work, but the *Convergence Claim* is. In fact, in the first part of this work it has been showed, through Sidgwick's arguments, how Intuitionism and Utilitarianism can be unified. In the second part we have seen how Parfit solved the Dualism of Practical Reason: he showed how there is no Dualism in morality, that the Practical Reason is unified, since Altruism *is* rational, and Egoism is not. In the third part it has been tried to demonstrate how it is possible to find a reliable *Theory X*, and how we need not to become moral skeptics if we carefully reflect on the relevant arguments. Becoming moral skeptics means believing that morality has no truths upon which everyone would agree: moral skepticism is probably the strongest denial of the *Convergence Claim*. In part three I showed how we need not to become moral skeptics. As a whole, this work shows that Practical Reason is unified, and that through reasoning an universal moral agreement can be reached. This work supports the *Convergence Claim.*

The *Convergence Claim* is the first connection between Parfit and Nietzsche's Aphorism §343. Parfit's and Nietzsche's widely free views are similar in

[424] Even if Parfit tries explicitly to prove this claim in all the volumes of *On What Matters*, it is defined in this form for the first time only in (Parfit, 2017, p. 309).

their enthusiasm for the possible progress of morality freed by religion, despite, to some extent, their view is different if we consider how the revision of Morality has to be made.

The second connection is humanity's new horizon. Nietzsche believed that humanity will be surpassed by a more perfect being, the *Übermensch*, the supra-human. Parfit does not describe the feature of the new humanity and hardly ever writes about his idea of new humanity, but he never denies that he hopes for a supra-human. This hope implies a great responsibility for those who live. At the end of *Reasons and Persons*, Parfit writes:

> I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
> (1) Peace.
> (2) A nuclear war that kills 99% of the world's existing population.
> (3) A nuclear war that kills 100%.
> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is *very much* greater.
>
> […] The Earth will remain inhabitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.[425]

It might be argued that he concluded this reasoning thirty-three years later, in the final pages of the third volume of *On What Matters*, the last book he published before his death. In the following words, the echoes of Nietzsche are clear:

> What now matters most is how we respond to various risks to the survival of humanity. We are creating some of these risks, and we are discovering how we could respond to these and other risks. If we reduce these risks, and humanity survives the next few centuries, our descendants or successors could end these risks by spreading through this galaxy.
>
> Life can be wonderful as well as terrible, and we shall increasingly have the power to make life good. Since human history may be only just beginning, we can expect that future

[425] (Parfit, 1984, p. 453-545)

> humans, or supra-humans, may achieve some great goods that we cannot now even imagine. In Nietzsche's words, there has never been such a new dawn and clear horizon, and such an open sea.[426]

Population Ethics is the branch of Ethics that seems a fit compass for venturing towards the clear horizon. Population Ethics, in fact, should provide the principles according to which humanity has to respond to the risks menacing it. If an universal agreement upon those principles will be reached and enough people will act according to them, humanity will be able to reach its apex – and so perhaps the universe.

> If we are the only rational beings in the Universe, as some recent evidence suggests, it matters even more whether we shall have descendants or successors during the billions of years in which that would be possible. Some of our successors might live lives and create worlds that, though failing to justify past suffering, would have given us all, including those who suffered most, reasons to be glad that the Universe exists.[427]

[426] (Parfit, 2017, p. 436)
[427] (Parfit, 2017, p. 437)

Bibliography

Arrhenius, G. (2000). Future Generations - A challenge for moral theory. *Dissertation for the Degree of Doctor of Philosophy in Practical Philosophy presented at Uppsala University in 2000.* Sweden, Uppsala.

Arrhenius, G. (2003). The Very Repugnant Conclusion. In K. S. Sliwinsky, *Logic, law, morality: thirteen essays on practical philosphy in honour of Lennart Aqvist* (p. 167-180). Uppsala: Departement of Philosophy.

Arrhenius, G. (2009). *One More Axiological Impossibility Theorem* (Offprint from Logic, Ethics and All That Jazz. Essays in Honour of Jordan Howard Sobel. ed.). Uppsala: Uppsala: Department of Philosophy, Uppsala University.

Arrhenius, G. (2011). The Impossibility of a Satisfactory Population Ethics. In H. C. Dzhafarov, *Descriptive and Normative Approaches to Human Behavior, Advanced Series on Mathematical Psychology.* World Scientific Publishing Company.

Arrhenius, G. (2016). Population Ethics and Different-Number-Based Imprecision. *Theoria, Volume 82, Issue 2, May*.

Arrhenius, G., & Campbell, T. (Forthcoming). The Problem of Optimal Population Size. *In Marc Fleurbaey, e.d.,International Panel on Social Progress 1st Annual Report*.

Briggs, R. (2014, August 8). *Normative Theories of Rational Choice: Expected Utility*. Tratto da Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/rationality-normative-utility/#Eth

Broome, J. (2006). *Weighing Lives.* Oxford: Oxford University Press.

Dasgupta, P. (2005). Regarding Optimum Population. *The Journal of Political Philosophy*, 414–442.

Dawkins, R. (2006). *The Selfish Gene.* Oxford: Oxford University Press.

Fehige, C. (1998). A Pareto Principle for Possible People . In U. W. Christoph Fehige, *Preferences* (p. 508-543). Berlin, New York: de Gruyter.

Grimes, W. (2017, January 4). *Derek Parfit, Philosopher Who Explored Identity and Moral Choice, Dies at 74.* Tratto da The New York Times: https://www.nytimes.com/2017/01/04/world/derek-parfit-philosopher-who-explored-identity-and-moral-choice-dies-at-74.html

Huemer, M. (2008). In Defence of Repugnance. *Mind, 117 (468)*, 899–933.

Hume, D. (1740). *Trattato sulla Natura Umana (Treatise on Human Nature)* (Vol. 1). (A. Carlini, A cura di, & A. Carlini, Trad.) Laterza.

Hume, D. (1751). Ricerca sui Principi della Morale. In D. Hume, & M. D. Pra (A cura di), *Ricerche sull'intelletto umano e sui principi della morale* (M. D. Pra, Trad., 1974 ed., p. 213-407). Bari: Biblioteca filosofica Laterza.

Justin, W. (2017, January 2). *Derek Parfit (1942-2017)* . Tratto da Daily Nous: http://dailynous.com/2017/01/02/derek-parfit-1942-2017/

Kwon, D. (2016, April 28). *The Battle over Pain in the Brain.* Tratto da Scientific American: https://www.scientificamerican.com/article/the-battle-over-pain-in-the-brain/

MacFarquhar, L. (2011, September 5). *How to be good.* Tratto da The New Yorker: http://www.newyorker.com/magazine/2011/09/05/how-to-be-good

Matthews, D. (2017, January 3). *The whole philosophy community is mourning Derek Parfit. Here's why he mattered.* Tratto da Vox: http://www.vox.com/science-and-health/2017/1/3/14148208/derek-parfit-rip-obit

Mill, J. S. (1858). *Saggio sulla libertà* (2009 ed.). (S. Magistretti, Trad.) Milano: Il Saggiatore.

Mill, J. S. (1861). *Utilitarianism* (2003 ed.). Oxford: Blackwell Publishing Ldt.

Nagel, T. (1971, May). Brain Bisection and the Unity of Consciousness. *Synthese 22*, 396-413.

Nagel, T. (1978). *The Possibility of Altruism.* Princeton, New Jersey: Princeton University Press.

Nakano-Okuno, M. (2011). *Sidgwick and Contemporary Utilitarianism.* New York: Palgrave MacMillan.

Ng, Y.-K. (1989). What should we do about future generations? *Economics and Philosophy, 5*, 235-253.

Nozick, R. (1974). *Anarchy, State, and Utopia.* Oxford: Blackwell Publishers Ltd .

obituary, (2017, January 4). *Derek Parfit.* Tratto da The Times: http://www.thetimes.co.uk/edition/register/derek-parfit-xvlxn0zq5

O'Grady, J. (2017, January 12). *Derek Parfit obituary.* Tratto da The Guardian: https://www.theguardian.com/world/2017/jan/12/derek-parfit-obituary

Parfit, D. (1971). Personal Identity. *The Philosophical Review, Vol. 80, N. 1*, 3-27.

Parfit, D. (1984). *Reasons and Persons.* Oxford: Oxford University Press.

Parfit, D. (2004). Overpopulation and the Quality of Life. In J. Ryberg, & T. Tännsjö , *The Repugnant Conclusion* (p. 7-22). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Parfit, D. (2011 a). *On What Matters* (Vol. 1). Oxford: Oxford University Press.

Parfit, D. (2011 b). *On What Matters* (Vol. 2). Oxford: Oxford University Press.

Parfit, D. (2016). Can we avoid The Repugnant Conclusion? *Theoria*, 110-127.

Parfit, D. (2017). *On What Matters* (Vol. 3). Oxford: Oxford University Press.

Rachlin, H. (2010, July). *How should we behave? A review of Reasons and Persons by Derek Parfit.* Tratto il giorno August 30, 2017 da National Center for Bioechnology Information: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893620/

Rawls, J. (1981). *Foreword to Henry Sidgwick' s Methods of Ethics, unabriged and unaltered republication of the 7th edition (1907) as published by McMillan and Company, Limited, Londo.* Indianapolis: Hackett Publishing Company.

Royal Swedish Academy of Science, K. O. (2014, January 13). *The Rolf Schock Prizes 2014: Rolf Schock – uniting philosophy, mathematics, music and art.* Tratto da Kung. Svetenskapsakademien: http://www.kva.se/en/pressroom/Press-releases-2014/The-Rolf-Schock-Prizes-2014/

Ryberg, J. (2004). The Repugnant Conclusion and Worthwile Living. In T. T. Jesper Ryberg, *The Repugnant Conclusion Essays on Population Ethics* (p. 239-255). Dordrecht, The Nederlands: Kluwer Academic Publishers.

Sidgwick, H. (1874). *Methods of Ethics* (unabridged and unaltered republication of the 7th edition (1907) as published by Macmillan and Company, Limited, London; with foreword by John Rawls ed.). Indianapolis: Hackett Publishing Company.

Sidgwick, H. (2000). *Essays on Ethics and Method.* (M. Singer, A cura di) Oxford: Oxford University Press.

Sperry, R. W. (1966). Brain Bisection and Mechanisms of Consciousness. In J. C. Eccles, *Brain and Conscious Experience* (p. 298-313). Berlin: Springer Verlag.

Tännsjö, T. (2004). Why we ought to accept the Repugnant Conclusion. In T. Tännsjö , & J. Ryberg, *The Repugnant Conclusion: essays on population*

ethics (p. 219-237). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Temkin, L. (2012). *Rethinking the Good. Moral Ideas and the Nature of Practical Reasoning.* Oxford: Oxford University Press.

Wittgenstein, L. (1953). *Philosophical Investigations.* (G. E. Anscombe, Trad.) Oxford: Basil Blackwell Ltd.