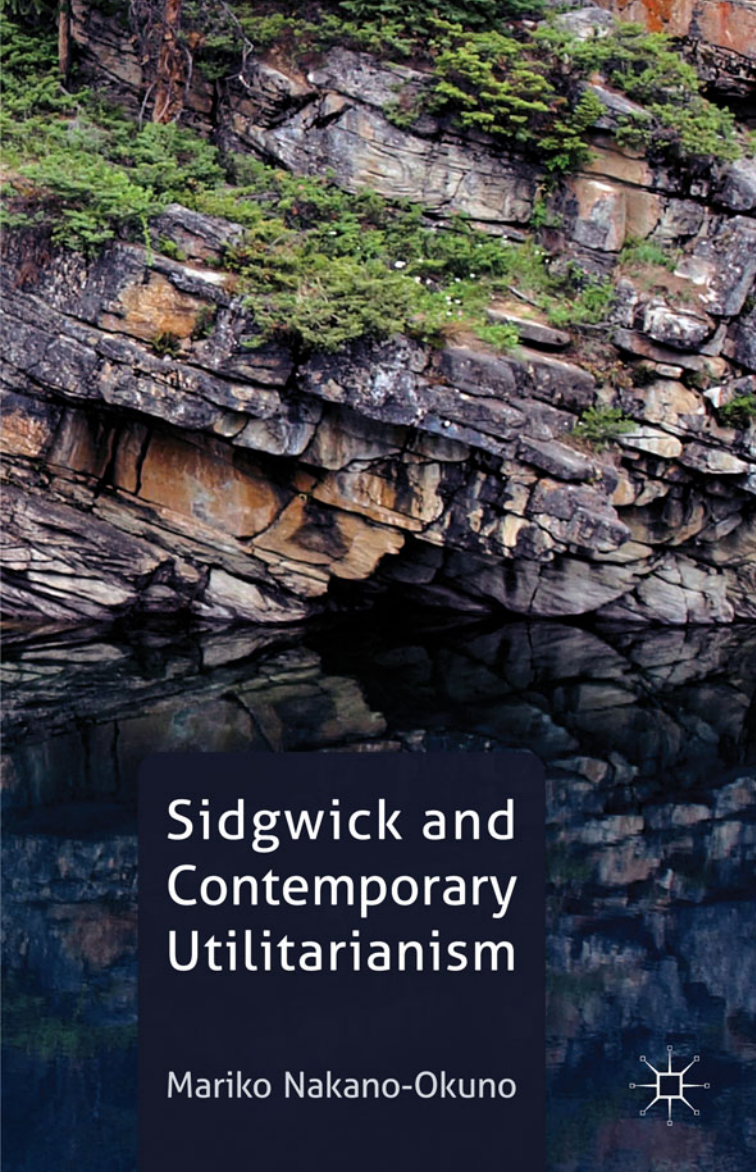

Sidgwick and Contemporary Utilitarianism

Mariko Nakano-Okuno

Sidgwick and Contemporary Utilitarianism

This page intentionally left blank

Sidgwick and Contemporary Utilitarianism

Mariko Nakano-Okuno
The Ohio State University, USA

palgrave
macmillan

First published 2011 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

ISBN 978–0–230–32178–6

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Nakano-Okuno, Mariko, 1970–
[Sidgwick to gendai kourishugi. English]
Sidgwick and contemporary utilitarianism / Mariko Nakano-Okuno
p. cm.
Includes bibliographical references (p.) and index.
ISBN 978–0–230–32178–6 (alk. paper)
1. Sidgwick, Henry, 1838–1900. 2. Utilitarianism. I. Title.
B1649.S44N3613 2011
170.92—dc23 2011021110

10 9 8 7 6 5 4 3 2 1
20 19 18 17 16 15 14 13 12 11

Printed and bound in Great Britain by
CPI Antony Rowe, Chippenham and Eastbourne

Table of Contents

List of Figures

Preface and Acknowledgements

This book deals with the moral philosophy of the nineteenth-century British utilitarian philosopher, Henry Sidgwick. However, the main theme of this book is not historical but rather an analysis of his impacts on contemporary ethics. Though this book focuses on the implications of Sidgwick's ethical theory to contemporary utilitarianism, this analysis extends beyond utilitarianism to include a more holistic examination of the philosophical foundations of ethics. This is done in light of an accurate interpretation of what John Rawls and Derek Parfit acknowledge as the best book in ethics: *The Methods of Ethics* written by Henry Sidgwick.

To be honest, Sidgwick's works did not appeal to me in the earliest stage of my career as an ethics researcher. I was heavily impacted by R. M. Hare's moral philosophy together with his two-level preference-utilitarian theory, and thus my research interest lay almost exclusively with contemporary works, such as those of Hare, Derek Parfit and R. B. Brandt. At that time, I could not see the point of investigating an old-fashioned scholar who lived a hundred years ago. Only when I noticed that both Parfit and Brandt, and even Rawls, who is considered to be one of the strongest opponents of utilitarianism, quite frequently refer to Sidgwick in developing their own arguments, did I finally start my in-depth study of Sidgwick. Might he have made a significant contribution to the field of contemporary ethics? Perhaps. However, it seemed to me that these contemporary authors had not fully explicated Sidgwick's own ethical theory. Rawls surely mentions Sidgwick when he criticizes classical utilitarianism in his major work (Rawls 1971, Ch. 1 Sec. 5), but his explanation of it seemed to contain some common misunderstandings of utilitarian ethics. Parfit develops in-depth discussion of philosophical topics by often referring to several passages from Sidgwick, but Parfit's own argument is so unique (while being important on its own) that we cannot discern Sidgwick's ethical theory by simply reading Parfit's works. The same is true for Brandt. R. M. Hare, the most influential utilitarian philosopher in the twentieth century, only briefly refers to Sidgwick. This led me to seek out several commentaries and anthologies about Sidgwick, including J. C. Schneewind's *Sidgwick's Ethics and Victorian Moral Philosophy* (Schneewind 1977) and Bart Schultz and others' *Essays on Henry Sidgwick* (Schultz 1992). I learned a

lot from them, but still did not understand how and in what regards Sidgwick's classical theory could be influential in contemporary ethics. Thus, I became determined to conduct my own study on Sidgwick *in the context of contemporary ethics*, with a focus on contemporary utilitarian ethics, which resulted in the publication of this volume.

Interestingly, the initial outline of my book turned out to be drastically different from my final draft. My early understanding of the history of utilitarianism was quite a conventional one; I believed that contemporary utilitarian thinkers had 'overcome' all or most of the theoretical difficulties with which Sidgwick left us. Thus, I sought to show exactly what theoretical problems Sidgwick left unresolved, and how we conquered them. However, once I started to read *The Methods of Ethics* and closely compare its arguments with those of contemporary utilitarians, I gradually came to realize that Sidgwick's insights were more profound, accurate and convincing than those authors. Thus, at a certain point, I had to entirely rewrite my draft, and to change my claim from 'we surpassed Sidgwick' to 'none of us can match Sidgwick'. I believe that this realignment helped me to highlight the significance of discussing Sidgwick today.

I will refrain from summarizing the content of this book here. Any comments, suggestions or criticisms are greatly appreciated, as they will encourage me to ponder the issues in a more profound manner.

Many people helped me in writing this book. This volume is a revised English version of my book published in Japan in 1999, which evolved from my PhD thesis written in 1998. I thank Professor Emeritus Hisatake Kato, who mentored me while I was a PhD student in the ethics department at Kyoto University, Japan, and advised me to study Sidgwick. Professor Kato facilitated the Japanese translation of *The Methods of Ethics* and invited me to participate as a co-translator. Thus, I started my study of Sidgwick in a quite favorable environment, in which I had many peers who were familiar with Sidgwick's works. Professor Emeritus Soshichi Uchii, who was my chief mentor until I obtained my master's degree and who later became Professor of Philosophy and History of Science at Kyoto University, continued to be my most influential advisor and always swiftly responded to all the questions and issues I addressed to him. The readers of this book will see the obvious influences of his remarkable analytical skills on my ways of thinking. Professor Emeritus Shigeru Yukiyasu at Okayama University helped identify the most important literature related to Sidgwick. I also thank Professor Masahiko Mizutani and my fellow researchers – including, but not limited to, Hideyuki Yahata, Shiro Shirouzu, Satoshi Eguchi, Nobuo Kurata, Tetsuji

Iseda, Makoto Suzuki, Taro Okuda, Satoshi Kodama, Yoshinori Hayashi, and Taku Sasaki – in the Department of Ethics at Kyoto University for discussing relevant topics with me and giving me useful comments. My special thanks also go to Masaru Tomioka, commissioning editor at the Keiso Shobo publishing company, who unconditionally agreed to publish the Japanese version of my book. I am also thankful to Susumu Morimura, Takeshi Ohba and Takashi Kawamoto, who gave me critical yet constructive comments on the Japanese version of this book. A year later the Japan Society of Ethics presented me with an award for this work, for which I am grateful. I must also thank the Japan Society for the Promotion of Science for financially supporting me from 1995 to 2000 through its grants and fellowships.

After teaching at Japanese universities for seven years, I moved to the United States with my family in 2005. In order to establish my academic credentials, I began translating my previously published papers and books, but I could not predict whether anyone would recognize me as an independent scholar in moral philosophy. In this context, I deeply thank Professors Utpal Banerjee and Luisa Iruela-Arispe in the Department of Molecular, Cell and Developmental Biology at the University of California, Los Angeles, for reading two of my draft papers – one on Kantian and utilitarian ethics and another on research on human embryos – and hiring me as a lecturer to teach biomedical ethics as a general education course at UCLA. I worked with a number of wonderful students there. I also deeply thank Professor Harley Kornblum, Drs Pamela Hurley, and Heather Tarleton at UCLA for all the training they gave me. Other intellectuals helped me through personal advice, conversation, seminars, and more. Those people include, but are not limited to, Professors Russell Korobkin, Linda and Edward McCabe, Seana Shifflin, Barbara Harman, Andrew Sabl, and Louis M. Guenin. Recently I started a new career at the Ohio State University in Columbus, Ohio, and I sincerely appreciate Dr. Catherine R. Lucey at the OSU Medical Center and Professor Donald C. Hubin in the Department of Philosophy for welcoming me as a new addition to the university. Thanks should also go to Drs Michael G. Bissell, Bruce Biagi, Pamela Salsberry, William Gardner, Piers Norris Turner, Carson Reider, Carol Hasbrouck, Lori Martensen, and many other faculty members and staff at the College of Medicine, the Department of Philosophy and the Center for Ethics and Human Values.

Also to be acknowledged here is the international Sidgwick research community, which consists of passionate scholars such as Professors and Drs Placido Bucolo, Bart Shultz, Roger Crisp, Philip Schofield, and

Hortense Geninet, among others. Though I have not had many chances to meet with them, there is no doubt that I have been influenced by their writings and activities, and, in some cases, through personal communication with them.

I wish I could list all the names of people to whom I would like to show my gratitude, but that is an impossible task to be met in this limited volume. However, there are seven people whom I cannot leave out. One is Professor Peter Singer at Princeton University. Since we met a decade ago, when he gave talks both in Tokyo and in Kyoto, Japan, Professor Singer continuously encouraged me to continue research and inspired me in many aspects by his own works. I deeply thank him for discussing crucial topics in ethics with me, and for reading the unfinished draft of this book.

Another person for whom I would like to show my deepest gratitude and respect is the late Professor R. M. Hare. When I met him in the summer of 1995, I asked him if he knew of any academic works that he would recommend to the students of contemporary moral philosophy, and he suggested several works, including Mane Hajdin's 1990 article which later became the main subject of Chapter 9 of the present book. I sincerely thank Professor Hare for letting me know the importance of Hajdin's paper, and hope that he would, were he alive, have taken my 'criticisms' of some parts of his arguments *not* as any sort of reproach against him, but as an honest effort to inquire into the nature of morality which both of us have eagerly pursued. I also thank Dr Mane Hajdin at Santa Clara University for immediately sending me the copy of his paper when I contacted him during the preparation of this English version. Obviously, my present volume would not exist had I not encountered these philosophers' works.

The fourth person whom I greatly appreciate is Kristina Nordstrom, a journalist and a great ESL instructor who proofread the entire draft with me. She not only corrected my expressions but also critically examined some of my arguments out of pure curiosity and intelligence, and discussed with me many relevant examples using her rich imagination and experiences. I will surely miss the time I spent with Ms Nordstrom. I also thank Kristin Smock for proofreading part of my final draft after I moved into Columbus, Ohio.

Last but by no means least, I wish to thank Priyanka Gibbons and Melanie Blair at Palgrave Macmillan for their steady support throughout the process of publishing this volume. Without them, this work would not have reached western readers. I also thank Michiko Doi and the Keiso Shobo Publishing Company for granting permission to publish

this English version from Palgrave Macmillan. I sincerely thank all the people I mentioned above, though I am the sole person responsible for all the claims addressed in the present book.

Finally, I thank my husband, Ichiro Nakano, and my daughter and son, Mayu and Yujiro, for always supporting me with their warm affections, showing me their vivid human feelings and teaching me how sympathy and imagination work together to generate an unimaginable power of benevolence.

Mariko Nakano-Okuno
Columbus, Ohio
February 2011

Notes on Abbreviations Used in the Text

1. Sidgwick, Henry (1907) *The Methods of Ethics*, seventh edition is abbreviated as *ME* or *ME*7. I mainly used the Hackett version published in 1981, a nonrevised reprint of the original version. When I refer to other editions of *The Methods of Ethics*, I use the abbreviation *ME* with the number of that edition. For example, *ME*1 refers to the first edition of *The Methods of Ethics*.

2. When I refer to other parts of this book, I use such descriptions as 'Chapter 1, Section 1, Subsection 1 of the present book', or '1.1.1'. On the other hand, '*ME* Book I, Chapter 1, Section 1' or '*ME* Bk. 1 Ch. 1 Sec. 1' means that part of *ME* (seventh edition, if there is no additional information).

3. Three other literary works are abbreviated as below:

FR = Hare, R. M. (1963) *Freedom and Reason*.
LM = Hare, R. M. (1961) *The Language of Morals*.
MT = Hare, R. M. (1981) *Moral Thinking*.

As for other works, please see the bibliography at the end of this book.

This page intentionally left blank

Introduction

When deciding what we ought to do, ordinary people do not always use one single method of ethical reasoning. They usually use various ways of ethical thinking that are addressed by different ethical theories. One sometimes acts from a sense of duty, sometimes thinks like a utilitarian, and sometimes behaves as if one is an egoist. A moral philosopher who never ignored this fact, and who explored the foundation of ethics by using his profound philosophical insight into our common moral thinking, was Henry Sidgwick (1838–1900), a nineteenth-century British utilitarian.

I have two aims in this book. First, I will elucidate the content of Sidgwick's theory of ethics, including his arguments for the philosophical foundations of utilitarianism. Second, I will reexamine, in the light of Sidgwick's theory, contemporary utilitarian theories, such as those of R. M. Hare, J. C. Harsanyi, and R. B. Brandt.

At this point, it might be helpful to give a rough idea of ethics and utilitarianism to be explored in this book. We are mainly concerned with 'ethics', taken as a study that explores the ways to reflect on what an individual ought to do when his or her action is expected to affect other people. This book will present utilitarianism as a way of ethical thinking in that sense.

As I see it, utilitarianism consists of several basic elements. The core of utilitarianism is the idea that a right action is one that will bring about maximum good on the whole for affected parties. (By 'affected parties' we mean all those who will be affected by the act in question. Some utilitarians consider not only human beings but also all sentient beings as affected parties, but I will henceforth use the simple term 'people' to denote both.) This idea can also be explained as a right action is the one that will bring about *consequences* which realize people's *maximum* good.

First, utilitarianism is a version of *consequentialism* or *teleology* in that it considers consequences or ends of an action to determine the rightness of that act. Second, it holds a kind of *maximization principle* in that utilitarianism evaluates an act in accordance with its tendency to bring about maximum good for all. Third, according to classical utilitarians such as Bentham, J. S. Mill, and Sidgwick, 'people's good' is construed as people's *happiness* or *pleasure* (a hedonistic theory of value). On the other hand, in a contemporary version of utilitarianism such as R. M. Hare's, it is explained as people's *desire- or preference-satisfaction* (a preference-based theory of value). Both classical and contemporary versions of utilitarianism interpret values in empirical terms, and insist that we should take into moral consideration such values as can be empirically identified by observation or introspection. This claim has been considered as one of the strongest points of utilitarianism. (There have been a few exceptions, as in the case of G. E. Moore's ideal utilitarianism, which claimed that the right act is one that will maximize good, and that good cannot be defined by natural properties such as pleasure or happiness. But I believe such nonempirical versions of utilitarianism are rather atypical and have never been supported for a long time.) Fourth and finally, people's pleasures or preference-satisfaction 'on the whole' means, according to many established utilitarians including Sidgwick, the *aggregation* of individual pleasures or preference-satisfaction. This point shows that utilitarianism takes a kind of individualistic view of public good, which claims that public good is reducible to the good of individuals. It should be noted here that many recognized utilitarians, including Sidgwick and Hare, adopt *total* utilitarianism that seeks to maximize the *sum total* of people's pleasures or satisfaction, whereas some people (like anti-utilitarian John Rawls, for example) proposed *average* utilitarianism as the most credible form of utilitarianism, in which one seeks to increase the average of people's pleasure or satisfaction.

This book will closely examine these fundamental elements of utilitarianism. It will elucidate where and how each of these elements – namely, consequentialism, the maximization principle, the hedonistic or preference-based theory of good, and the idea of aggregation – is derived from, and how those elements get combined to build up a utilitarian ethical theory. Through such investigation, I will present the reasons why utilitarians recommend the utilitarian way of ethical thinking. Additionally, I will consider the reason why we have morality in the first place.

I assume that most of us attempt to do some kind of ethical thinking. One, we often try to justify our behavior. Two, we often criticize others

that they are acting unethically, and three, we sometimes engage in public debates on some ethical issue of social interest. In order to better deal with these matters, some turn to ethics for help. But, what kind of help do we seek from ethics? As for myself, I sometimes cannot easily decide what I ought to do. In such cases I wish to obtain a clear, consistent, and convincing way to determine what is morally right or wrong, so that I no longer need to struggle with perplexing ethical dilemmas or moral conflicts. I suppose many others have a more or less similar wish, judging from the fact that many people care about ethical dilemmas and conflicts. The demand for clarity, consistency, and persuasiveness becomes even more serious when we have to make a public moral decision, in which we cannot decide what one ought to do just by asking ourselves what we *want* to do. Some questions about what to do may be solved just by clarifying one's own desire or wish, but when it comes to a moral conflict between persons or an ethical dilemma that involves other people, we have to make the effort to find a solution that can be widely accepted. But when we debate over such public ethical issues, we sometimes get confused and our discussion meets an impasse. The role of ethics is to get rid of such confusion or blocks in our moral thoughts and arguments, by sorting out problems and guiding us to a systematic, rational way of ethical reasoning based on a solid philosophical foundation.

Utilitarianism was proposed as a theory that offers us such a clear and consistent way of ethical thinking. It has had its detractors, who claimed that it leads us to a counterintuitive conclusion, or that its theoretical basis is flawed, but utilitarianism has survived and has been supported by contemporary thinkers such as Hare, Brandt, and Singer. Its survival owes a lot, I think, to Sidgwick's elaborate work, *The Methods of Ethics*, which is known as one of the best books on ethics, full of careful analyses and deep insights about the philosophical basis of utilitarianism. I will attempt to reevaluate the foundation of utilitarian ethics, following the tracks left by one of the greatest moral philosophers, by closely examining *The Methods of Ethics* (the seventh edition, henceforth abbreviated as *ME*).

We should keep in mind the definition of utilitarianism stated above, for it might be quite different from what people commonly understand as utilitarianism. First of all, utilitarianism should not be identified with egoism. I will later argue (11.2.2) that a selfish person could also come to adopt utilitarian ethics, but egoistic and utilitarian ways of thinking are two different types of logical reasoning. Utilitarianism pursues people's total happiness or satisfaction, whereas egoism pursues individual

pleasure or satisfaction. Second, what is meant by happiness or pleasure is not a mere material or economic interest, but, as will be later explained, any kind of feelings that satisfy or gratify the person feeling them. Third, utilitarians do not always pursue convenience or expediency. It is certain that utilitarianism might recommend a more convenient or effective method to obtain pleasure if such a method brings about an equal amount of pleasure compared with alternatives. If such a method diminishes the amount of resulting pleasure, however, utilitarians will by no means approve it. Fourth, though utilitarianism is a kind of consequentialism, it is not a theory which retrospectively evaluates past actions by looking into their results. Utilitarianism primarily considers what our future acts ought to be. Therefore, utilitarianism is basically prospective thinking, and it considers consequences that are *expected* at the time of our decision-making.

Here it would be proper to articulate why it is so important to study Sidgwick in a contemporary context. Henry Sidgwick is regarded as the last classic utilitarian and the pioneer of contemporary moral philosophy (see, for example, Albee 1901, p. 358; Rawls's foreword to the Hackett version of *ME*7 1981, pp. v–vi). Indeed, Sidgwick is one of the key figures in the history of utilitarian ethics, in that he maintained the main idea of classic utilitarianism and undertook the task of answering the early criticism against classic utilitarianism. His unique methodology in developing deeper arguments on the foundation of ethics has led to today's analytical philosophy. His impartial, levelheaded, and minute arguments, and his use of detailed analyses and criticisms, are highly reputed (Broad 1930, p. 143; Schneewind 1977, p. 1; Shionoya 1984, pp. 137–8). So, the advantage of studying Sidgwick seems obvious. By appropriating his analytical skills, we can reexamine the basic concepts of ethics and deepen our understanding of ethics and utilitarianism.

But the merits of studying Sidgwick are more than that. One remarkable characteristic of Sidgwick was the comprehensiveness and neutrality of his arguments. Though he was a utilitarian, in *The Methods of Ethics*, Sidgwick committed himself to strictly indifferent analyses of multiple methods that are used in our daily moral thinking – that is, utilitarianism, egoism, and a kind of deontology. Actually, and quite interestingly, Sidgwick often critically analyzed utilitarianism and pointed out its theoretical weakness. He even admitted that he had left some problems unsolved regarding the theory and method of utilitarianism. For example, Sidgwick plainly suggested that utilitarianism has difficulty with measurement and interpersonal comparisons of pleasure and pain. He also conceded that his proof of ethical hedonism, which

was needed to support his version of utilitarianism, cannot be conclusive. Moreover, at the end of the concluding chapter of *The Methods of Ethics* he confesses that he cannot perfectly reconcile utilitarianism with egoism. I will explore how contemporary utilitarian thinkers approach the questions that Sidgwick left open, and to what extent they have successfully solved them. When we examine such contemporary arguments in the light of Sidgwick's ethics, however, we will come to recognize several crucial points that are overlooked by contemporary writers, despite the advances they have made. Sidgwick's analysis was deeper than their analyses in some respects, and his point of view can still be utilized for criticizing these contemporary writers.

Thus, this book consists of two parts. Part I deals with Sidgwick's ethical theory. In Part II, I reexamine contemporary utilitarian arguments, based on my interpretation of Sidgwick's moral philosophy.

The aim of Part I is to precisely interpret Sidgwick's ethical theory, and to elucidate the structure of the theoretical foundation of utilitarianism. I will make clear – even more clearly than Sidgwick did – the analysis of the basic concepts of ethics, and the content and the meaning of the fundamental principles of ethics presented by Sidgwick. Then I will explicate Sidgwick's proof of hedonism, and the mechanism of deriving utilitarian total-maximization.

In Part II, I will focus on four theoretical difficulties that Sidgwick left, and I will examine how contemporary utilitarian thinkers have dealt with those problems. Contemporary works can often be appreciated as having further developed Sidgwick's analyses. At some points, however, I will rather reevaluate Sidgwick, pointing out some faults of contemporary arguments by utilizing discoveries we have made through our examination of Sidgwick's ethics. If I can successfully show that, despite their achievements, contemporary utilitarians have not yet reached Sidgwick's level, the significance of studying Sidgwick today will become sufficiently obvious.

The following four points are the major features of this book:

1. This volume spells out the exact meaning of each of Sidgwick's three fundamental principles of ethics, namely, the Principles of Justice, Prudence, and Benevolence. Furthermore, by introducing a new interpretation of these principles, referred to as 'Independent Interpretation' in Chapter 6, this book clarifies and emphasizes the essential differences among these principles, especially the important distinctions between the Principle of Justice and the Principles of Prudence and of Benevolence.[1]

2. It carefully reevaluates Sidgwick's proof of hedonism through an examination of Mane Hajdin's debates on Hare's preference-utilitarianism.
3. It closely analyzes the derivation of the maximization principle of utilitarianism. After clarifying how Sidgwick's own argument can be reconstructed, we will see how contemporary writers such as Arrow and Harsanyi have contributed to deepening the analysis.
4. Based on a full-fledged interpretation of Sidgwick's theory, this book critically examines the defects and weaknesses of contemporary utilitarian theories, such as those of Hare and Brandt.

Only a few contemporary researchers fully digested, clearly explained, and thoroughly examined Sidgwick's analysis and proof. Additionally, it is quite rare for researchers to apply his elaborate analysis to a reconsideration of contemporary utilitarianism.[2] The final goal of the present book is to clearly show that the above four points will highlight the significance of studying Sidgwick today.

Part I
Sidgwick's Theory of Ethics

Part I deals with Sidgwick's theory of ethics. In Chapters 1 and 2, I make an exegesis of Sidgwick's view on the scope of ethics, clarify the aim and the basic structure of *The Methods of Ethics*, and also posit the status of utilitarianism in this book. In Chapters 3 to 6, I examine his analyses and his theoretical claims in more detail, and in Chapter 7, I show how his analyses and claims have laid the foundation of utilitarianism.

This page intentionally left blank

1
The Scope of Ethics

1.1 What ethics is

Ethics as Sidgwick understands it is 'the science or study of what is right or what ought to be, so far as this depends upon the voluntary action of individuals' (*ME* p. 4). Science here means a systematic study, which seeks to attain precise knowledge (see *ME* p. 1). But unlike such sciences as psychology or sociology, ethics deals not with mere facts, but with norms of action, expressed in terms of 'ought' or 'right'. Some believe that ethics also deals with virtues, or a moral evaluation of a person's character, rather than his or her actions; but one's character is known to us only through his or her acts; and we usually do not separate an evaluation of someone's character from that of his or her behavior (*ME* Bk. 1 Ch. 9 p. 113 fn. 1). Therefore, the primary subject of ethics is considered to be an individual's actions. Another study that treats norms is politics; but unlike politics, which deals with the decision-making of a government, ethics deals with the actions of an individual. Thus Sidgwick's primary concern is with individual decision-making, but he does not distinguish between individual acts done in public and those done in private.[1]

We should note here that Sidgwick thinks the scope of ethics includes not only 'moral' action in the narrow sense – that is, moral action that solely considers altruistic consideration or duty to others – but also every action considered as 'the action that ought to be done', whether prudent or benevolent. As I will explain later, a rational egoist, who acts prudently to seek his own happiness, can also use the term 'ought'.

1.2 'Action' as a subject matter of ethics

1.2.1 Voluntary action

Ethics deals with the actions of an individual insofar as they are voluntary ones. 'Voluntary action' is, firstly, a conscious act that is distinguished from an unconscious or automatic bodily movement. Strictly speaking, it is an act being done while the agent is conscious of his act and of the self who is doing it. Secondly, Sidgwick says that voluntary action that is dealt with in ethics is not just a conscious act but an act that is consciously done with some *intention*. When one intends to do something, one is recognizing, or representing in one's mind, the consequences or effects that will be brought about by such an intention or by an action caused by such an intention. In addition, voluntary action is not an impulsive act, in which one's sentiment and action are directly connected, and it is accompanied by the consciousness that one is choosing the intended consequences. In short, ethics considers 'voluntary action' to be an act in which the agent is conscious of that act's consequences and of himself who is choosing and determining those consequences (*ME* Bk. 1 Ch. 5 Sec. 2 pp. 59–61).

1.2.2 Distinction between will and desire-or-motive

Sidgwick distinguishes will from desire or motive (*ME* p. 363 fn. 1). Roughly speaking, among impulses toward some object or action, the ones which an agent is aware of are called *desires*. Those desires directed toward an action, or toward the expected consequences of an action, are called *motives*. On the other hand, *will* makes a conscious choice among motives and decides to perform a single action.

More accurately, 'desires' are what one feels as impulses that urge a person to aim at a certain object, or as impulses or stimuli that urge a person to perform an act that has the tendency to obtain a certain object (*ME* p. 43 fn. 2). When this object is expected to be obtained as a consequence of an action, the desire for that consequence is called a 'motive'. Motives are desires, or conscious impulses, for consequences that will be realized by actions (see *ME* pp. 202, 362). Other types of impulses include instinctive or subconscious ones where the agent is not conscious of the targeted object, or of his act to realize the object.

Every action is done out of a certain impulse, but we often have multiple impulses simultaneously. When one's will does not work, one naturally acts according to the strongest impulse. But when one makes a conscious choice among those impulses to lead oneself to a certain act, it is said that one's volition is involved. In such a case, each desire or

motive *stimulates* one to will a certain action, but it is not one's desire or motive but one's will that finally decides which action one takes.

An act chosen by will is, however, not always performed, because this volition is sometimes overcome by other impulses. Again, from the fact that an act was chosen by one's will, it does not follow that this act is rational or right. Such a voluntary choice may turn out to be wrong, and there may be some other 'right' act that could have been willfully chosen. Ethics deals with voluntary actions that are consciously chosen by the agents' will, but the rightness of such actions is yet to be examined.

1.2.3 A note on 'consequences'

In the previous sections I explained that ethics deals with voluntary or intentional action and that every intentional action contains an internal recognition of certain consequences of that intention or intended action. Here we should present some points regarding the notion of 'consequences'. First, we should clarify what this denotes. Among the effects that can be caused by one's intention are (a) changes in the external world resulting from muscular movement, (b) changes in ideas and feelings that constitute our conscious life, and (c) changes in one's tendencies to act in certain ways under certain circumstances (see *ME* pp. 72–3). Sidgwick claims that in moral or legal discussions it would be best for us to take into consideration, under the term 'intention', *all* the consequences that are foreseen as certain or probable (*ME* p. 202) – though we should reconfirm that what is meant here is the expected effects and not those unforeseen ones which actually happened (*ME* p. 201).

We also have to explain here the distinction between deontology and consequentialism. Acts according to duty, or acts out of the sense of duty, are also regarded as a subject of ethics. It is commonly said, however, that acts according to duty are somewhat different from those done with a certain end or with an intention to bring about a certain consequence. Still, we can say that even the acts done from the sense of duty are done, insofar as they are intentional, with the realization of the consequence that 'an act according to duty has been done'. What distinguishes consequentialism from deontology is whether the act is done with the recognition of *other* consequences than that. It is such other consequences that Sidgwick meant by the term 'ulterior consequences' (*ME* pp. 8, 98, 170, among others). Consequentialism, utilitarianism being a typical example, regards such *ulterior* consequences as essential for judging the rightness of actions. We can distinguish between deontology and consequentialism because we commonly draw

a line between an act and its ulterior consequences at the point when the agent's purposeful bodily movement is completed, even though such movement completion can be seen, in a sense, as a part of the consequences of the agent's intention. Suppose, for instance, I intend to deceive someone by (1) telling him 'this is a real picture drawn by Renoir'. On being told so, (2) he may hold the mistaken belief that this picture is real, and by holding such a belief (3) he might purchase that picture at an exorbitant price. We can say that (1) (2) and (3) are, in a sense, the foreseeable consequences of my will to tell him a lie. But we usually include (1) in the act of telling a lie and regard (2) and (3) as ulterior consequences following my act of lying. According to deontology, one should foresee (1) and judge that one ought not to do this because it is an act of lying, whereas consequentialists judge that this act is wrong because the subsequent effects such as (2) and (3) are bad (*ME* Bk. 1 Ch. 8 Sec. 1; Bk. 3 Ch. 1 p. 200 fn. 3).

1.2.4 Matters that are not Sidgwick's chief concern: The goodness of motives, the formal or subjective rightness of acts

Getting back to our previous point, ethics is primarily concerned with the rightness of an individual voluntary action, which is chosen by the will of that individual and done with his or her intention to do it while foreseeing some of its consequences.

> Moralists of all schools, I conceive, would agree that the moral judgments which we pass on actions relate primarily to intentional actions regarded as intentional [. . .] When I speak therefore of acts, I must be understood to mean – unless the contrary is stated – acts presumed to be intentional and judged as such.
>
> (*ME* pp. 201–2)

Some may argue, however, that we evaluate our acts by looking at the *motives* of the agent. As we saw before, motives are desires or conscious impulses toward consequences realized by actions. The notion of a motive is distinct from that of an intention. For one thing, the essence of a motive is an impulse, whereas the basic meaning of an intention is foreseeing the consequences. For another thing, the consequences that are desired in one's motive are only part of the miscellaneous consequences that are foreseen in one's intention. It is often said that, as in the case of committing perjury in order to save the lives of one's parents or loved ones, one's motive can be judged as good even if one's intention is wrong. What is meant here is that though this person did

wrong in that he knowingly committed that act while foreseeing the result of committing perjury, we cannot regard his act as totally wrong because his desire to bring about the consequence of saving lives was a good one. According to Sidgwick, however, the rightness or wrongness of an act, or what ought to be done, is to be judged not by one's motives but by one's intention, and not by mere desired consequences but by all the foreseeable consequences – though he admits that motives are often taken into consideration, to some extent, when we evaluate someone's actions (*ME* pp. 202, 204).

We may still persist in a commonly accepted idea that evaluation of the agent's motive or desire to conform to duty plays an important role in our moral judgment. One may express such an idea by saying that one's action is 'formally' right if the agent is moved by pure desire to fulfill duty for duty's sake and not by desire to bring about certain ulterior consequences (*ME* pp. 206–7). However, when we consider what one ought to do in the near future, we are asking what one should intend to do, rather than whether one has a motive to fulfill whatever one thinks is his duty. What we would like to know in such a case is not the formal rightness of action, which depends on the agent's desire to act according to duty, but the substantial (or, to use Sidgwick's term, 'material') rightness of an action, which is judged by the various effects the agent will bring about.

Again, an action may be called 'subjectively' right, when the agent performs that act because he *believes* it is his duty, even though he may not have *a desire* to fulfill his duty. Sidgwick insists, however, that ethics should inquire into what is called the 'objective' rightness, rather than the subjective rightness of an act (*ME* pp. 207–8). Certainly, many people think that an action cannot be absolutely right when the agent believes it to be wrong. In addition, we are often impressed with a person who acts with a firm belief that he is doing the right thing, and we respect the subjective rightness of his will unless it causes enormous harm to someone. Nevertheless, for those who turn to ethics, being unable to determine what is the right thing to do, it would be paradoxical to answer that the right thing to do is to do what one believes to be right. Such an answer does not seem to offer further systematic development, nor is it a useful guide for us. In addition, admitting that one cannot distinguish between the subjective and objective rightness of one's own act, we frequently judge other people's actions to be subjectively right but objectively wrong. For example, we often consider a fanatic's act wrong. Under the name of ethics, we deal with actions done by the agent's own will, but the rightness of such actions must be judged on

grounds other than a mere self-conviction of the agent himself. When an action is judged to be right, it should be either because its ends are objectively right (that is, such ends can be regarded as right not only by the agent but also by other people), or because the reason for such an act is objectively right.

To sum up, whether the agent's motive is good, whether the agent is motivated by pure desire of performing duty, or whether he is motivated by the belief that it is his duty is not the primary concern of ethics in Sidgwick's sense.

1.3 What a method of ethics is

So, ethics is a study of how to judge the objective rightness of the intentional acts of individuals. But where can we start such a study?

According to Sidgwick, we have two ways to investigate ethics: one is to look for the true moral laws or rational precepts of conduct and the other is an inquiry into the nature of good as an ultimate end (*ME* pp. 2–3). Sidgwick adopts the former, since he thinks that we inquire into the nature of good in order to obtain a guide for our actions. We turn to ethics not just to gain an understanding of what good is, but to decide what we ought to do, or to decide what is the right thing to do, based on such an understanding. Another reason why Sidgwick adopts the former approach is because some believe that the right act is what is unconditionally ordered as duty, without assuming any good which can be attained by that act. Sidgwick thus begins his inquiry not by reviewing various opinions about the nature of good, but by collecting various views about the laws or precepts that are supposed to guide our actions – though it will later turn out (in 4.3 of the present book) that we still need to analyze the concept of good.

Further, Sidgwick doesn't merely clarify the contents of common laws or precepts of conduct, but investigates the *methods* of guiding an action according to such laws or precepts of conduct. We seek such precepts or laws in order to guide our conduct, but they are of no use unless we can clarify how to decide our actions based on those laws and precepts. A 'method of Ethics' is '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' (*ME* p. 1). A *rational* procedure is considered to have some consistent line of reasoning, expressed in the form of a principle that states such precepts or laws for guiding actions.[2] And as a *procedure*, it is supposed to show us a concrete process in which we can make a rational decision of what ought to be done.

But we commonly use not one but at least three methods of ethics. There are three systematic methods based on three different views, namely, egoism, utilitarianism, and dogmatic intuitionism. To put it simply, these three views can be described as follows:

1. ***Egoism, or Egoistic Hedonism***: The ultimate end of one's action should be one's own happiness or pleasure, and one ought to perform an act that will best accomplish this end.
2. ***Utilitarianism, or Universalistic Hedonism***: The ultimate end of one's action should be other people's happiness or pleasure as well as one's own, and one ought to perform an act that will best accomplish this end.
3. ***Dogmatic Intuitionism***: One ought to act according to some moral rules (dogmas) that are apprehended by our intuition.

These views state three different practical principles about what one ought to do, all of which we are ready to accept as apparently valid and legitimate. The methods of ethics that are logically derived from these principles are called, respectively, the method of egoism, the method of utilitarianism, and the method of dogmatic intuitionism.

2
An Overview of *The Methods of Ethics*

2.1 The background

2.1.1 Sidgwick's concern as a utilitarian

It is evident that Henry Sidgwick was a utilitarian, from his writings and from the testimony of people closest to him. For example, in *The Elements of Politics*, Sidgwick asserts that there is a general assent among people that the ultimate criterion for determining right and wrong in legislation is a utilitarian one (Sidgwick 1891, Ch. 3 Sec. 2 pp. 34–5; 3rd edn, 1908, pp. 37–8). In his 1897 essay, he also states that 'for those who, like myself, hold that the only true basis for morality is a utilitarian basis' ('Public Morality', in Sidgwick 1898, p. 63).

What was remarkable about Sidgwick, however, was the fact that in his *Methods of Ethics* Sidgwick did not manifestly advocate utilitarianism. Instead, he simply examined three methods of ethics – that is, those of egoism, dogmatic intuitionism, and utilitarianism – equally, as the methods we commonly use, and which Sidgwick himself used in his moral thinking.

The reason why Sidgwick took that line of argument is stated in an autobiographical manuscript included in the preface to the 6th edition of *ME*, which was published after his death. Such is often mentioned by several writers (Uchii 1988, p. 216 ff.; Schneewind 1977, p. 40 ff. and others). According to that manuscript, Sidgwick first devoted himself to J. S. Mill's utilitarian ethics. Sidgwick confesses that by getting to know Mill's utilitarianism he found the way to be released from the bridle of conventional morality, which was often confusing and dogmatic. Then, Sidgwick read the writings of William Whewell (1794–1866), one of the most popular scholars of his time as well as an 'intuitionist' moral philosopher, who advocated the authority of precepts or rules of

common-sense morality. Sidgwick, however, had the impression that the definitions and axioms in Whewell's arguments were quite lax and imprecise. Thus, Sidgwick decided to rigorously analyze the basic moral concepts and principles, to critically examine the method of dogmatic intuitionism which claims the authority of common moral rules, and to compare it with the method of utilitarianism. This became the first theme of Sidgwick's moral philosophy.

Sidgwick gradually came to have some doubts on Mill's argument for utilitarianism. To Sidgwick, Mill seemed to derive the utilitarian ethical claim that everyone ought to aim at general happiness for all affected parties from the psychological fact that everyone does seek one's own happiness; but Sidgwick came to consider this argument as philosophically insufficient. Doesn't the fact that everyone actually seeks one's own happiness support the claim of egoism that everyone ought to aim at one's own happiness, rather than the utilitarian claim? Further, isn't it undeniable that everyone believes not only that one ought to aim at people's happiness, but also that it is reasonable to seek for one's own happiness? Becoming aware of these questions, Sidgwick started to examine the method of egoism as an independent ethical view from utilitarianism. This is the reason why the method of egoism is also discussed in *The Methods of Ethics*.

But if Mill's 'proof' of deriving utilitarianism from the psychological fact that everyone does seek for one's own happiness is wrong, on what philosophical grounds can we support utilitarianism? Sidgwick slowly came to accept the idea that such grounds can be obtained only by our fundamental intuition, though such an intuition should be a fully reflective, refined and philosophical one. As he launched his inquiry into the fundamental intuition(s) of ethics, he found in the writings of Immanuel Kant one such fundamental intuition which Sidgwick held himself. Sidgwick also noticed that Joseph Butler, whom Sidgwick considered a mere advocate for the authority of conscience, admitted the rationality of self-love. Furthermore, Sidgwick discovered in the writings of early intuitionists, such as Henry More (1614–87) and Samuel Clarke (1675–1729), the original form of the axiom necessary to establish utilitarianism. Having realized that he and the other predominant philosophers, including intuitionist philosophers, shared the same fundamental intuitions, Sidgwick decided to step into an Aristotelian examination to reflect on various types of current common moral opinions in an impartial manner.[1] The style of Sidgwick's argument has thus been established. In *The Methods of Ethics*, Sidgwick first makes rigorous analyses of the basic moral concepts, and then examines three methods of ethics (egoism, dogmatic

intuitionism, and utilitarianism), while searching for the fundamental intuitions that would become the foundations of those three methods by referring to our commonsense.

If we glance at the ideological trend at his time, we can further understand the reason for his style of arguing. From the late eighteenth to the nineteenth century, William Paley (1743–1805), Bentham (1748–1832), and J. S. Mill (1806–73) proposed utilitarianism as the first principle of morality and the only valid method of ethical decision-making which should supersede all the existing traditional teachings and moral sentiments. Utilitarianism was frequently criticized by intellectuals such as Samuel Taylor Coleridge (1772–1834), Whewell, and John Grote (1813–66). Some of them attacked Mill's proof of utilitarianism, and others refused the assumption that everyone is basically selfish and always seeks one's own happiness. Still others denied the idea that the notion of duty is reducible to that of interest. Such opponents usually took the position that assumed the existence of our moral faculty or conscience, and proposed multiple moral principles such as those of benevolence, justice, sincerity, veracity, purity, fidelity, and so on. At the same time, however, thinkers such as Whewell and Grote also admitted that utilitarianism contained a partial truth, and even the Christian moralists of the time commonly thought that the notion of human good or welfare must contain happiness as well as the fulfillment of one's duty (*ME* p. 3). But it was unclear how they absorbed what utilitarianism ordered them to do, when it often conflicted with what their common-sense morality told them to do. Also unclear were their criteria for making decisions on moral conflicts, in which multiple moral principles contradict each other. In the end, they managed to remove such worries about possible conflict, by assuming that God would bring about overall harmony (for a detailed discussion of this point, see Schneewind 1977, Chapters 2 and 3, especially pp. 66, 100, 105–6, 111).

Since the main subject of this book is to discuss the significance of Sidgwick's argument in a contemporary context, I will not go further into the history of British ethical thought. My main concern is not with how Sidgwick's thought developed against the ideological background of his time, but with what Sidgwick can provide to the contemporary problems we face. But what I would like to point out through the above description is that Sidgwick, while seriously espousing utilitarianism, equally seriously considered the criticisms of utilitarianism. He never ignored the fact that we certainly have nonutilitarian ways of thinking. Being aware of that fact, he undertook the task of reconciling utilitarianism with non- or anti-utilitarian ways of ethical thinking, such as

moral views that put much weight on our conscience, common-sense morality or moral sense.

Sidgwick is not the only person who focused on the seemingly irreconcilable relationship between utilitarianism and common-sense morality, or between utilitarianism and duty or virtue theories (see, for example, Bellah et al. 1996; Sandel 2009 among others). The task that Sidgwick undertook is what we still wish to accomplish in today's context.

2.1.2 The aim and fundamental assumption of *The Methods of Ethics*

In our ordinary moral thinking, we certainly use any of the three methods of ethics. Sidgwick was well aware of this. But the problem is that we often use these methods in a confused way, and that we are often troubled with the question of what one really ought to do – or, more precisely, by which method one should determine what one ought to do. We wish to eliminate or diminish this uncertainty and confusion in our thought. Hence, as the primary aim of *The Methods of Ethics*, Sidgwick set himself the task of defining multiple methods of ethics that can be found in our common-sense morality, explicating the consequences of adopting each of those methods, and elucidating how these methods are related to each other in our moral thinking.

> My object, then, in the present work, is to expound as clearly and as fully as my limits will allow the different methods of Ethics that I find implicit in our common moral reasoning; to point out their mutual relations; and where they seem to conflict, to define the issue as much as possible.
>
> (*ME* Bk. 1 Ch. 1 Sec. 5 p. 14)

As I repeatedly point out, egoism, utilitarianism, and intuitionism are the three ways of thinking in an individual's moral reasoning, rather than three different ethical positions, each of which has its own advocates. Ethics attempts to systematize our thoughts regarding what we ought to do and keep them consistent. Consequently, *The Methods of Ethics* seeks to clarify the contents of these three methods that are commonly confused, and to understand their mutual relationship as clearly and coherently as possible.

When we undertake such a study, we are already making one fundamental assumption. This is the assumption that our reasoning as to what one ought to do can become a consistent and harmonious one,

even though under the present circumstances the three methods often conflict, or are often confused. Therefore, Sidgwick proceeds with the assumption that 'We cannot regard as valid reasonings that lead to conflicting conclusions' and that 'so far as two methods conflict, one or other of them must be modified or rejected' (*ME* p. 6). Such an assumption is called a fundamental postulate of ethics.

2.1.3 How to read *The Methods of Ethics*

Can we, then, regard *The Methods of Ethics* as the work that defended utilitarianism, because it was written by a utilitarian? Did Sidgwick, by clarifying the contents and the mutual relationship of these three methods of ethics, show the supremacy of utilitarianism over the other two methods?

The Method of Ethics first examines egoism and intuitionism, and then utilitarianism (*ME* Bk. 4), which certainly gives us the impression that for Sidgwick utilitarianism was the most important among the three ways of moral thinking. In addition, in some passages, Sidgwick apparently commits to utilitarianism. Furthermore, readers of the first edition must have understood this book as fully defending utilitarianism because of the structure and contents of its overall arguments.[2]

However, in the preface of the first edition of *ME*, Sidgwick proclaims that this work aims to expound and examine the methods of ethics from a neutral point of view. He plans to examine what conclusions we may reach when we adopt a certain method, which assumptions we make, and how accurately we come to those conclusions. While doing this analysis, he claims to not favor one of these methods. Additionally, in the preface of the second edition, Sidgwick deplores that it is a misunderstanding that some people regarded the first edition as defending utilitarianism while attacking the methods of intuitionism and egoism. In the main text of *The Methods of Ethics*, he manifestly states as follows:

> In the course of this endeavour [to expound the multiple methods of ethics] I am led to discuss the considerations which should, in my opinion, be decisive in determining the adoption of ethical first principles: but it is not my primary aim to establish such principles [. . .]. I have wished to keep the reader's attention throughout directed to the processes rather than the results of ethical thought: and have therefore never stated as my own any positive practical conclusions unless by way of illustration: and have never ventured to decide dogmatically any controverted points, except where the controversy

> seemed to arise from want of precision or clearness in the definition of principles, or want of consistency in reasoning.
>
> (*ME* Bk. 1 Ch. 1 p. 14)

The Methods of Ethics simply intends to critically explain the three methods that ordinary people use in their moral thinking – that is, to define them, to elucidate their contents, and to point out their mutual relationships. It is not the main purpose of *ME* to establish the higher principle that integrates those methods. Besides, Sidgwick honestly admits in the concluding chapter of *ME* that he could not prove the supremacy of the utilitarian method in his book. Therefore, though Sidgwick was a utilitarian, *The Methods of Ethics* itself is not the work to refute the other two methods of ethics,[3] nor to establish utilitarianism as the only legitimate moral view.[4]

Despite all this, in my opinion it does not mean that Sidgwick had no wish to defend utilitarianism by writing *The Methods of Ethics*. What he wanted to avoid was to presuppose the rightness of one single method and to reject the others. Sidgwick never meant that it was a mistake to use the other two methods, nor that we should abandon those two. Rather, he was trying to show that, though we will continue to use all three methods in our ordinary moral thinking, it is the utilitarian method that can settle conflicts or ambiguities that could occur at times. He presumably hoped that, by showing this, he could come to claim that utilitarianism is the theory that ultimately governs all methods of ethics.

> To eliminate or reduce this indefiniteness and confusion is the sole immediate end that I have proposed to myself in the present work. In order better to execute this task, I have refrained from expressly attempting any such complete and final solution of the chief ethical difficulties and controversies as would convert this exposition of various methods into the development of a harmonious system. At the same time I hope to afford aid towards the construction of such a system; because it seems easier to judge of the mutual relations and conflicting claims of different modes of thought, after an impartial and rigorous investigation of the conclusions to which they logically lead.
>
> (*ME* p. 13)

Sidgwick certainly wished to offer aid to construct a harmonious ethical system. It is highly probable that he hoped such a system to be a

utilitarian one. However, in order to construct a truly reliable, harmonious ethical system, it is quite important not to defy the actual diversity of our common moral thinking. Thus, Sidgwick confined the aim of *The Methods of Ethics* to neutral analyses of the multiple methods of ethics that we commonly use, to clarification of their contents and limitations, and to elucidation of their mutual relationships. Such clarifications are *prerequisites* for establishing the primacy of the method of utilitarianism. Sidgwick states that 'We cannot but hope that all methods may ultimately coincide: and at any rate, before making our election we may reasonably wish to have the completest possible knowledge of each' (*ME* p. 14). My understanding is that *The Methods of Ethics* is a work that developed impartial and elaborate arguments as a preliminary preparation for advocating utilitarianism. In those arguments, Sidgwick, while keeping his neutral stance, certainly presented a path to support utilitarianism.

Therefore, although Sidgwick declared that he would remain neutral in *The Methods of Ethics*, and although *ME* was not meant to prove utilitarianism, I will regard this book as providing sufficient resources to establish the consistent ethical system behind classic utilitarianism, as well as that of contemporary utilitarianism.[5]

2.2 The overall structure of *The Methods of Ethics*

The Methods of Ethics consists of four books plus a concluding chapter, and their arguments can be summarized as follows.

Book I

In Book I of *ME*, the scope of ethics is strictly stipulated, and egoism, utilitarianism, and dogmatic intuitionism are identified as the three methods of ethics. At the same time, basic moral concepts such as 'ought', 'right', 'pleasure', and 'good' are closely analyzed. Such conceptual analysis becomes the most important basis for the entire arguments in *The Methods of Ethics*.

Book II

In Books II to IV, the three methods are examined in turn. Book II deals with egoism. It details the method of egoism, and examines the practical procedure to choose actions that will best promote one's own pleasure. Sidgwick points out that several such procedures are proposed. (1) According to the 'empirical-reflective' method, one should foresee one's own pleasure and pain resulting from each alternative action as

precisely as possible, then measure and compare the amount of such pleasure and pain, and finally choose one action that will bring about the greatest amount of one's happiness. (2) The 'objective' method of egoism tells us to ascertain not feelings of pleasure and pain, which are difficult to measure, but particular things that are commonly judged to bring about such feelings. For example, social status, fame, power, etc. are regarded as things that will produce pleasure. The objective method urges us to choose actions that will bring about such results. (3) The 'deductive' method means that one should decide which action will bring about the greatest pleasure, based on deductions from general principles about the sources of pleasure and pain. Such general principles proposed in Sidgwick's time include George Frederick Stout's suggestion that pleasures are connected to moderate activities as opposed to excessive or insufficient activities of the body, and Herbert Spencer's claim that pleasures are the correlatives of acts tending to continue or extend one's life. After examining these three methods, Sidgwick supports the empirical-reflective method of assessing pleasure and pain. He argues that, though the empirical-reflective method involves the complexities and difficulties of hedonistic calculation, our knowledge of the resources or general principles of happiness used in the objective or deductive method is even more uncertain and ineffective.

Book III

Book III investigates the method of dogmatic intuitionism, which claims that we can intuitively judge the rightness or wrongness of actions and that common-sense morality reflects our intuition. Since dogmatic intuitionism usually relies on the general rules of common-sense morality (duties and virtues), Sidgwick makes one-by-one examinations of particular duties and virtues. By doing so, he shows that every rule has some limitations and exceptions, and that the boundaries of right and wrong set by such rules prove to be vague on close examination. Sidgwick further points out that, when two or more rules come into conflict, common-sense morality cannot provide a consistent guide for determining what we ought to do. Thus, the limits of dogmatic intuitionism become clear.

The most important argument in Book III appears in Chapter 11, which comes after the critical discussion of common moral rules. In Chapter 11 of *ME*, Sidgwick presents four (or, actually, five) necessary conditions for truly self-evident and significant propositions. After explaining these conditions, Sidgwick reconfirms that particular rules of common-sense morality do not meet these conditions and hence do

not qualify as truly self-evident valid standards for our moral behavior. However, since these five conditions are always to be presupposed when we search for self-evident and significant propositions, we should assume that 'the three, truly self-evident, fundamental principles', which are presented later, must satisfy these conditions.

After clarifying the limits of dogmatic intuitionism through such examination, Sidgwick departs from the investigation of dogmatic intuitionism and starts to search for more sophisticated, abstract moral principles, using what he calls philosophical intuition. In Chapter 13 of Book III, Sidgwick discovers three such 'truly self-evident and significant' fundamental principles. It is these three principles that make up the core of Sidgwick's ethical theory. These principles are expressed in several ways, but they can be described as follows:

1. *The Principle of Justice*: If someone's particular act is right for that person, that act is right for every similar person in similar circumstances.
2. *The Principle of Rational Self-Love*: One ought to aim at one's own overall good. In doing so, one ought to give equal weight to one's good at one moment and to one's good at another moment, unless there is a difference in amount or certainty of the good that one expects to obtain.
3. *The Principle of Rational Benevolence*: One ought to aim at people's overall good, as long as one takes a universal point of view that regards one's own good as a part of people's overall good. In aiming at such people's overall good, everyone is morally obliged to place equal weight on one's own good and on the good of others, unless he judges that another's good is lesser in amount, or that another's good is less certain to be precisely known or obtained.

Sidgwick argues that these are the principles that can be found in our reflective commonsense. Meanwhile, of these three principles, the Principle of Rational Benevolence can be construed as equivalent to what is known as Bentham's dictum, 'everybody to count for one, and nobody for more than one'. Sidgwick thus claims that this principle provides one of the philosophical foundations of utilitarianism. In the following chapter (*ME* Ch. 14), Sidgwick attempts to demonstrate 'proof of hedonism', in order to show that the 'good' used in his fundamental principles is ultimately comprised of nothing but pleasure. This proof of hedonism, coupled with the fundamental principles, especially that of Rational Benevolence, provides the theoretical basis of utilitarianism.

Book IV

Finally, utilitarianism is closely examined in Book IV. The meaning of utilitarianism is fully explained, and its relationship to common-sense morality is discussed. In Sidgwick's opinion, ordinary people are following the moral rules of common-sense morality without recognizing themselves as utilitarians, but they implicitly determine the details of applications of such rules in a utilitarian way (the 'unconscious utilitarian' hypothesis). Then, Sidgwick reviews particular virtues and duties, and points out that the limitations and exceptions of these rules are determined by utilitarian considerations. Sidgwick further claims that, when it is unclear whether one should apply a certain rule to a given situation, or when such rules conflict with each other, we implicitly solve problems in a utilitarian way – or, at least we *could* solve such problems by applying the method of utilitarianism. Thus, utilitarianism, the theoretical basis of which was given in Book III, is further proved to be useful in systematizing our moral thinking.

The latter half of Book IV explains how we can utilize the utilitarian method in practice. Sidgwick claims that the most cogent method is, again, the empirical-reflective one, which foresees people's actual pleasure and pain, and then measures and compares their amounts, and finally chooses the alternative that will bring about the greatest happiness for all. However, it is impossible for humans to make such calculations and to establish brand new rules on each occasion. Sidgwick argues that, for us, the most appropriate way of applying utilitarianism is to generally respect existing moral rules while making suitable modifications and revisions of them. Thus, common moral rules serve as 'middle axioms' which mediate between the utilitarian principle and a concrete guide for actions, but such rules are founded by the utilitarian principle and revised according to that principle. The only feasible method the utilitarian principle can endorse is, Sidgwick claims, the empirical-reflective one. However, Sidgwick admits that even in utilitarianism the empirical-reflective method necessitates certain assumptions about hedonistic calculations, and contains several areas of difficulty.

Concluding chapter

Based on the analyses and examinations in the previous chapters, the concluding chapter of *ME* discusses the mutual relationships of the three methods of ethics.

As for the relationship between dogmatic intuitionism and utilitarianism, Sidgwick concludes that they are reconcilable with each other. He argues that the rules of common-sense morality, which dogmatic

intuitionism supports, are to be generally observed even from the utilitarian point of view. However, when conflicts or ambiguities of such rules hinder us from making consistent moral judgments, we should appeal to utilitarianism to solve these problems. This means that the method of dogmatic intuitionism is after all dependent on, and controlled by, utilitarianism.

However, as for the relationship between egoism and utilitarianism, Sidgwick confesses that he cannot perfectly reconcile them. According to Sidgwick, egoism is not irrational nor irrefutable. He alleges that egoism has a solid theoretical foundation as does utilitarianism. The combination of the Principle of Rational Self-Love and the proof of hedonism establishes egoism, whereas the Principle of Rational Benevolence and the proof of hedonism establish utilitarianism – so the theoretical basis of egoism is quite on a par with that of utilitarianism. Furthermore, Sidgwick claims that, as far as we judge from our earthly experiences, we have to admit that there may be a conflict between what egoism tells us to do and what utilitarianism tells us to do, and that we cannot harmonize them unless we make a religious assumption that God will make them coincide by rewarding a benevolent person and punishing a selfish person in the afterlife. According to Sidgwick, when such a conflict of egoism and utilitarianism occurs, one's practical reason will be torn apart and become unable to decide what one ought to do. Thus, 'the dualism of practical reason', that is, the antithetical relationship between egoism and utilitarianism, is left unsolved in *The Methods of Ethics*.

2.3 The status of utilitarianism in *The Methods of Ethics*

In *The Methods of Ethics*, utilitarianism is seen from four different perspectives, that is, its place in our ordinary reasoning, its relation to common moral rules, its theoretical basis, and its practical method.

First, utilitarianism is a kind of moral reasoning that people commonly use. It is but one of the three ways of thinking about what one ought to do, and ordinary people also use the methods of egoism and dogmatic intuitionism. On reflection, however, it turns out that utilitarianism provides us with a more consistent guide than dogmatic intuitionism. But egoism and utilitarianism are equally consistent and reasonable, and we cannot systematically decide which line of reasoning we should follow when they order us to pursue different courses of action.

Second, it is utilitarianism that can make sense of, revise, and systematize common-sense morality. It is a misunderstanding that utilitarianism

runs counter to common-sense morality, because commonly accepted moral rules are generally significant even from the utilitarian point of view.

Third, utilitarianism is proved to have a well-defined theoretical basis, being supported by two philosophical intuitions, the Principle of Rational Benevolence and ethical hedonism.

Fourth, the practical method of utilitarianism necessitates the assumption that hedonistic calculation is possible, and such a calculation contains several practical difficulties. But egoism is not really superior to utilitarianism on this point, because a similar assumption and similar difficulties accompany egoism. On the other hand, dogmatic intuitionism, as far as it appeals to common moral rules, is to be ultimately governed by utilitarianism.

In short, in *The Methods of Ethics* utilitarianism is regarded as a way of thinking that (1) can validate and systematize common-sense morality, (2) has a well-built theoretical basis, (3) involves several difficulties in its decision-making procedure, and (4) cannot show its supremacy over egoism. As Sidgwick predicted in Chapter 1 of *The Methods of Ethics*, *ME* does not establish utilitarianism as a sole valid ethical theory, though its theoretical grounds and practical usefulness are suggested. What is remarkable about Sidgwick is that he, while acknowledging himself as a utilitarian, quite candidly points out the difficulties of utilitarian calculations of pleasure and pain, and even admits his 'failure' to reconcile utilitarianism and egoism. This shows nothing but honesty on the part of Sidgwick, who concentrates on neutral analyses of the three methods.

Then, what contribution did Sidgwick make toward contemporary utilitarian ethics? One such contribution would be his elaborate examination of the relationship between utilitarianism and common-sense morality, because contemporary critics of utilitarianism still often claim that utilitarianism is simply counterintuitive, or lacking commonsense. Actually, the greater part of *ME* is dedicated to an argument showing that utilitarianism regards commonly recognized moral rules as generally important, and that only utilitarianism can systematize common-sense morality. For contemporary ethics, however, what is no less important is that Sidgwick attempted to clarify the theoretical foundation of utilitarianism and suggested that such foundation consists of the Principle of Rational Benevolence and the proof of hedonism. We should also appreciate his analyses of basic moral concepts, such as 'good', 'right', and 'ought', which were presented as the basis of his entire arguments on ethics.

However, as repeatedly pointed out, Sidgwick has not shown that utilitarianism is the only valid ethical view. His last confession of the dualism of practical reason is a clear example of this. I will later explain Sidgwick's argument on the foundation of utilitarianism in more detail, and then reexamine the validity of utilitarian ethical theory, referring to recent philosophical arguments on relevant topics.

In the next chapter, we will look into the basics and the methods of egoism, utilitarianism, and dogmatic intuitionism in more detail than the preliminary explanation given above. We will next clarify the meanings of 'intuition' and 'commonsense', because these are two key notions that Sidgwick adopts as a starting point for his philosophical investigation into ethics. He also refers to his and other people's intuition and commonsense frequently, in order to make sure that his analyses or reasoning is truly convincing. Sidgwick uses the term 'intuition' and 'intuitionism' in several different senses and at different levels, however, so we should carefully distinguish among those different meanings.

Then we will examine the basic elements that ultimately constitute Sidgwick's claim about the philosophical foundation of utilitarianism. Such elements include (1) *analyses of basic moral concepts* such as 'right', 'ought', and 'good', (2) *five necessary conditions* for 'truly self-evident and significant propositions', (3) *three fundamental principles of ethics*, and (4) *the proof of hedonism*. Of them, Sidgwick actually presented (1), (2), and (3) as the basis of ethics in general, but they should also be regarded as crucial elements to construct the utilitarian theory. But utilitarianism cannot be established without (4), the proof of hedonism. We also have to show how two distinctive features of utilitarianism, *consequentialism* and *the total-maximization principle*, are logically derived. Furthermore, we have to fully understand Sidgwick's claim that only utilitarianism can systematize common-sense morality. This last claim is regarded as another support for the credibility of utilitarianism. The remaining part of Part I of this book will address these issues.

3
Three Methods, Intuition, and Commonsense

3.1 Three methods of ethics

3.1.1 Egoism

Egoism is the view that the ultimate end of an individual's acts is that person's happiness or pleasure, and that one ought to act so that one can accomplish this end (see, for example, *ME* Bk. 1 Ch. 7 p. 89 and Bk. 2 Ch. 1 p. 119). 'The ultimate end' means an end that should be sought in itself, that is, not as a means to any other ends. We should also note that the above view is a normative claim about what one *ought to* do, and not a psychological claim that one *does* always seek one's own happiness in his voluntary actions.

Here we should not regard egoism as *any* theory that refers to 'ego' or 'self'. If we literally interpreted 'egoism' as the view by which one judges what one ought to do based on a certain self-related principle, every ethical view would be included in the vast category of egoism, because it involves an individual's self-conscious actions. As opposed to such an over-comprehensive definition of egoism, Sidgwick restricts his meaning of egoism to the view that one ought to promote one's own *happiness* or *pleasure*. Hence, egoism is also called egoistic hedonism.

Happiness is interpreted as equal to pleasure, or as what is constituted by pleasures (*ME* p. 92). In brief, pleasure denotes the kinds of feelings which are desired by the agent himself, and which can be expressed as 'agreeable' or 'desirable' feelings. Likewise, pain denotes the 'undesirable' feelings that are disliked and averted by the agent. As I noted in 1.2 of the present book, desire means what one feels as an impulse urging one to attain some target, or an impulse urging one to perform an act to achieve that goal. When the target of desire is the agent's own feeling, that desired feeling is called pleasure. When one acts according to the

impulse or motive to realize that desired feeling, and when the desired feeling actually occurs as a result of such an act, we can say that pleasure has been generated.

I will later analyze the notion of pleasure, but here we should remember that Sidgwick uses the terms pleasure or happiness in quite a broad sense, to include any kind of pleasurable, agreeable, or satisfied feelings. The term 'happiness' is sometimes distinguished from definite specific pleasures such as the gratification of sensual appetite or other keen and vehement desires, and used to mean much calmer and more moderate contentment. However, Sidgwick would clearly call such calmer contentment 'pleasure' as well, for he includes *any* kind of agreeable feeling in the notion of pleasure (*ME* pp. 92–3).

The decision-making procedure logically derived from this view of egoism is the method of egoism. It naturally encourages one to decide what one ought to do by choosing an action that will directly or indirectly promote one's own happiness or pleasure. Specifically, egoism claims that the ultimate end of one's action is the *greatest* attainable happiness of one's own. I will closely analyze how this condition of 'greatest' comes about. I should point out at this time that Sidgwick never claims that an egoist should aim at his immediate pleasure or a lesser pleasure. Rather, he asserts that an egoist ought to choose an act that will bring about the greatest attainable pleasure for himself, by considering his entire situation, present and future. Every action is done according to a certain impulse, and pleasure or satisfaction of any kind results from accomplishing an act to gratify that impulse. Therefore, when one has multiple impulses to do particular acts that are mutually incompatible, the method of egoism tells a person to compare expected pleasures that will result from those impulses and actions, to identify an action that will bring about the greatest overall pleasure for himself, and to reinforce the impulse toward that act. The desire for one's own greatest attainable pleasure, or the desire for one's own pleasure in general, is said to be operating here, and it is called 'self-love' (*ME* pp. 89, 93).

Then, how can we determine which is the 'greater', or 'lesser', pleasure among those expected pleasures that one could experience? Such a measurement of the amount of pleasure or happiness is done in terms of such pleasure's 'desirability' or 'preferability', that is, the intensity of one's desire or preference toward the feeling of pleasure. In other words, the comparison of pleasures is made by determining which pleasure is more strongly preferred by the person himself, when he envisions all the expected pleasures precisely in his mind, while excluding all the factors outside of those feelings (I will fully explain this later). 'The

intensity of pleasure' means the intensity of a person's desire or preference toward his pleasure. Here, Sidgwick emphasizes that this 'intensity of pleasure' is *different from* the intensity of the *sensation* of a pleasure. Sidgwick claims that 'a pleasant feeling may be strong and absorbing, and yet not so pleasant as another that is more subtle and delicate' (*ME* Bk. 1 Ch. 7 p. 94). The intensity of pleasure depends not on how vividly and powerfully that sensation is felt by the person experiencing it, but on how intensely that person prefers that feeling. Sidgwick labels this comparison of pleasure based on the intensity of preferences as a *quantitative* comparison of pleasure. In other words, what is *more intensely* preferred in the above situation is the *greater amount of* pleasure. Sidgwick claims that the only viable method of comparing pleasures is this quantitative comparison just described. He does not admit the comparison of the *quality* of pleasures, which John Stuart Mill proposed. In Sidgwick's opinion, when one says that a certain pleasure is better in quality than the other – for example, when one says that a pleasure resulting from people's mutual affection and trust is superior to that resulting from satisfying one's appetite, what is meant is that the former type of pleasure is more pleasant than the latter (*ME* pp. 94–5). We often call an experience which is partly pleasant but contains pain, or which involves subsequent pain, an 'impure' or 'low-class' pleasure, but this actually means that the amount of pleasure, or the surplus of pleasure over pain included in the experience, is relatively less than that of other pleasant experiences. On the other hand, if we stick to the idea that the quality of pleasure is completely different from its quantity, we would have to judge such quality using criteria that appeals to something *other than pleasure*. Such a determination would not be pure egoistic hedonism, which considers one's own pleasure as the sole criteria in deciding what one ought to do. Rather, it should be regarded as a position indistinguishable from intuitionism, which will be later described. Thus, in egoism, the quality of pleasure is reduced to its quantity. 'The greatest attainable happiness' means the greatest attainable amount of happiness, and is defined as 'the greatest attainable surplus of pleasure over pain'. According to Sidgwick, the method of egoism is, and should be, described only in terms of the quantity of pleasure and pain. The method of egoism is a procedure which tells us to choose an action that is expected to bring about the maximum surplus of pleasure over pain.

> There remains then Pure or Quantitative Egoistic Hedonism [. . .]. According to this the rational agent regards quantity of consequent pleasure and pain to himself as alone important in choosing between

> alternatives of action; and seeks always the greatest attainable surplus of pleasure over pain – which, without violation of usage, we may designate as this 'greatest happiness.
>
> (*ME* Bk. 1 Ch. 7 p. 95)

One last point to note: egoism is a view that adopts one's own happiness as the *ultimate criteria* for judging what ought to be done, and it does not necessarily claim that one should always consciously seek one's own happiness when acting. Our experience shows that it is often true that a person can best attain his own happiness if he acts from other motives than a conscious pursuit of his own pleasure – to put it differently, if he acts according to a decision-making process that does not directly aim at his own happiness. If this is true, such a way of acting, or such a procedure, can be justified from the egoistic standpoint. Therefore, though I described the basic policy of the method of egoism above, we still have to examine multiple possibilities of its actual decision-making process, which can be practically adopted in the application of egoism.

3.1.2 Utilitarianism

Utilitarianism is the view that the ultimate end of one's action is the general happiness of mankind – that is, to maximize overall happiness or pleasure of all parties affected by that act – and that the right action is one that will best accomplish this end (*ME* pp. 8, 411–13).

> By Utilitarianism is here meant the ethical theory, that the conduct which, under any given circumstances, is objectively right, is that which will produce the greatest amount of happiness on the whole; that is, taking into account all whose happiness is affected by the conduct.
>
> (*ME* Bk. 4 Ch. 1 p. 411)

Another name for utilitarianism is universalistic hedonism. Sidgwick uses the term hedonism because utilitarianism also seeks happiness or pleasure as the ultimate end. However, unlike egoism that tells one to seek one's own happiness, utilitarianism tells us to seek the happiness of all parties (humans, or sentient beings) whose happiness will be affected by the act in question. Hence, it bears the adjective 'universalistic'. Sidgwick uses the term 'universal' in several contexts, but its basic meaning is 'being applicable not only to a single particular, but also to *anyone* or *anything* that belongs to a certain definable class' (see *ME* p. 34). The expression 'universal happiness or pleasure' is supposed to mean

the happiness or pleasure of all individuals who are sentient beings, and not merely of a single individual. We should notice that the term 'universal' can be used either in the sense that refers to any individual human or sentient being, or in a much wider sense that refers to any individual thing, which could be nonhuman or even nonanimal. Along with the expression 'universal happiness or pleasure', Sidgwick often uses the expression 'the general happiness of humankind'. 'General' here means 'overall' or 'including all members of humankind (or all sentient beings)'. However, we should note that the term 'general' can be used in several other senses. It sometimes means 'applying to *most* cases' or 'widely admitted', as when we say 'the general consensus of people', and it sometimes means 'by and large', 'normal', or 'basic', as when we say 'the general meaning of a term'. At times, this term may also have the connotation of 'not being specific'.[1] For instance, when we use phrases such as general principles or general notions, we envision rules or concepts that are described in nonspecific and broad terms. Whereas universality puts an emphasis on not singling out particular individuals within a certain definable class, generality gives us the impression that a set of people, things, or cases is grouped together in a category which can be described in simple, generalized terms. Additionally, 'general' does not necessarily mean that such a generalized or simplified description must be applied to all the members without exception. Anyway, we can say that what Sidgwick meant by the phrases 'universal happiness' and 'general happiness' is almost the same. Both mean people's overall happiness, as contrasted with an individual's happiness.

The decision-making process logically derived from a utilitarianism perspective is called the method of utilitarianism. Its basic policy is to choose an action that will maximize people's overall happiness. As with the method of egoism, Sidgwick claims that the comparison of pleasures should be done by assessing how intensely such pleasures would be desired or preferred (under certain conditions as explained earlier) by the persons feeling them. However, in utilitarianism we have to compare and maximize not a single individual's pleasure but people's overall pleasure. The greatest amount of general happiness is that in which the sum total of aggregate happiness of individuals is greater than any other possible alternatives. Such an amount is considered the surplus of the sum total of people's pleasure over the sum total of people's pain. If we consciously seek to make such a calculation, we have to compare, add, and subtract people's pleasure. This interpersonal comparison and aggregation of people's pleasure is a crucial step in the method of utilitarianism, which will be closely examined later in this book.

Again, utilitarianism is a view that adopts people's universal happiness as the *ultimate criteria* for judging what one ought to do, and it does not necessarily claim that one should always consciously seek such universal happiness. Also in the method of utilitarianism, we still need to consider several possibilities of practical decision-making. Such a process may recommend that we act without being conscious of the ultimate end at the time of our action (*ME* p. 413).

3.1.3 Dogmatic intuitionism

Dogmatic intuitionism is a kind of view that was supported by Sidgwick's contemporaries, who were called intuitionists. The word intuitionism can have other meanings, too, and different aspects of intuition will be explained later in this chapter. Here, I will mainly explain the most commonly known version of intuitionism.

What is usually known as intuitionism is, in Sidgwick's opinion, the view that 'a conduct is held to be right when conformed to certain precepts or principles of Duty, intuitively known to be unconditionally binding' (*ME* p. 3). 'Intuitively' in this citation can be construed as 'immediately and without using any inferences'. Duty is defined as the right action or nonaction that is imposed on someone, when a possible motive exists that urges one to do a wrong act, and when one needs a moral motive to defeat such immoral motives in order to perform the right act (*ME* p. 217). Intuitionism in this sense is a view that one can immediately apprehend precepts or rules that prescribe the right actions as unconditionally binding, and that one can judge the rightness or wrongness of actions in the light of those precepts or rules.

Sidgwick includes the view that one ought to be virtuous, or to act according to virtue, in this meaning of intuitionism. Virtue means the relatively permanent laudable qualities of a person manifested in conduct performed with voluntary effort. An excellence that can be immediately manifest without any effort of the will is called a gift, and not a virtue (see *ME* Bk. 3 Ch. 2 p. 220). Sidgwick claims that whether one has a virtuous character or not can be judged mostly by seeing whether one is acting according to the common rules of duty, or to certain precepts or rules that prescribe admirable conduct. Intuitionism is a way of thinking in which one intuitively apprehends such precepts or rules of conduct, and in which one judges how one ought to act according to those rules. This includes what are called deontology and virtue theory today.

According to this view, one can ascertain that an action is right (or ought to be done) just by knowing that the action is following a precept or a rule of duty or virtue. Related to this, moralists following intuitionism

in Sidgwick's time emphasized that this judgment of the rightness of an action should never depend on the ulterior consequences of the action in question. This is evidently a nonconsequentialist claim. As I suggested before, as long as ethics deals with intentional actions, and intention always involves some kind of consequence of that intention, we could say that even intuitionism deals with consequences. It surely considers the resulting situation in which a certain act has been done, as a consequence of the agent's decision. However, intuitionists are unwilling to allow for other (that is, *ulterior*) consequences when they determine whether an action is right or wrong. They tell us to judge the rightness of an action solely by considering whether the action is conforming to duty.

However, if intuitionists simply claim that they never take into account any other consequence than the one that an action has been done, their claim is incorrect for two reasons. First, some duties and virtues approved by the intuitionist school certainly consider the ulterior consequences of actions. For example, prudence and benevolence are generally regarded as virtues. When one considers one's own future or pursues other people's happiness in order to realize these virtues, one has to take into account various ulterior consequences of one's actions – namely, the action's direct and indirect effects on one's future happiness or on people's happiness. What distinguishes, then, intuitionists from egoists or utilitarians is that intuitionists consider such ulterior consequences insofar as they are suggested by the precepts or rules of duty. The point here is that intuitionists do not embrace other ends than the fulfillment of the precepts or rules of duty, and hence do not consider other consequences than the ones that are to be brought about when the duty is fulfilled. Second, as I suggested before, it is often difficult to draw a line between an act and its ulterior consequences. So, the claim that intuitionists do not think of any ulterior consequence applies only to those kinds of acts in which the ordinary usage of language allows us to draw a sufficiently clear line between acts and their ulterior consequences – for example, 'truth-telling' enables us to draw a relatively clear line between the act of telling the truth and its later consequences (*ME* Bk. 1 Ch. 8 Sec. 1).

Therefore, intuitionism, as it is commonly known, is the view that 'certain kinds of actions are right and reasonable in themselves, apart from their consequences; or rather with merely partial consideration of consequences, from which other consequences admitted to be possibly good or bad are definitely excluded' (*ME* Bk. 3 Ch. 1 p. 200).

According to this view, a person is required to ascertain precepts or rules of duty that are unconditionally prescribed, and to choose an act

which abides by such precepts or rules. This process of deciding what one ought to do is called the method of intuitionism. This procedure is different from the methods of egoism or utilitarianism in several respects. One such difference is that whereas the basic principle of egoism or utilitarianism, which is that one ought to promote one's own or people's happiness, can roughly determine what kind of action one ought to take, the basic principle of intuitionism, which is that one ought to act according to the precepts and rules of duty or virtue, is not enough for one to determine what one actually ought to do. It further requires a person to clarify the content of each particular precept or rule before he can actually decide what he ought to do. According to intuitionism, these right precepts or rules are intuitively apprehended. This view cannot hold without the assumption that we have a special ability to clearly apprehend truly valid precepts and rules of conduct.

At this point, it should be noted that so far I have always used two terms, 'precepts' and 'rules', to describe the idea of intuitionism. The common idea of intuitionism actually contains two distinct types of views. One claims that certain precepts, which prescribe particular actions, are intuitively known, and the other claims that certain general rules (generalized prescriptions that can be applied to multiple cases), which stipulate certain kinds of actions, are intuitively known. An example of the former type is such a view that one can judge by one's conscience the rightness of an action in each particular case. Sidgwick calls this type of view 'perceptional intuitionism'. However, perceptional intuitionism can only provide us with a decision-making procedure which is not very helpful, namely, that the agent instantly apprehends what he ought to do, or that one ought to act according to the judgment of a person whose conscience has a good reputation. We cannot expect further systematic development of perceptional intuitionism. Only the latter type of intuitionism, in which one apprehends certain general rules, is the view which can develop a systematic method of ethics, and which deserves philosophical investigation. Sidgwick calls this view 'dogmatic intuitionism', and in *ME* he mainly examines the method of dogmatic intuitionism as one of the three methods of ethics.

According to the method of dogmatic intuitionism, one determines what one ought to do by ascertaining the general rules of duty or virtue that one thinks one should obey unconditionally, and by asking if what one plans to do conforms to those rules. In order to systematize this method, we need to enumerate and spell out those rules of conduct. Such general rules are likely to be implied in commonly acknowledged duties and virtues. For example, widely accepted rules of duty such

as keeping a promise or not telling a lie, and virtues such as wisdom, benevolence, justice, and courage are promising candidates for the rules that dogmatic intuitionism approves. Therefore, one of the tasks in our examination of the method of dogmatic intuitionism is to precisely formulate the rules of common-sense morality, and to clearly understand what these rules order us to do.

The above is a basic explanation of the three views and methods of ethics. We can summarize their differences as follows. First, egoism and utilitarianism prescribe a certain act as a means to some other end which one ultimately ought to accomplish, whereas dogmatic intuitionism prescribes a certain act of duty or virtue, as what ought to be done in itself and not as a means to some other end. Second, egoism and utilitarianism both aim at pleasure or happiness, but they differ greatly in that egoism pursues an individual's self-regarding pleasure, whereas utilitarianism pursues people's universal pleasure.

3.1.4 Sources of the three methods

At this point, one might naturally ask how these methods of ethics are derived, and why all kinds of ethical thinking come down to these three.

Sidgwick starts his discussion to identify methods of ethics by contemplating the ultimate reasons for actions (see *ME* Bk. 1 Chs. 1 and 6).

> What then do we commonly regard as valid ultimate reasons for acting or abstaining? This, as was said, is the starting-point for the discussions of the present treatise.
>
> (*ME* p. 78)

A method of ethics is a rational procedure by which one determines a particular act to be performed. For it to be a rational one, it cannot be such a procedure in which one randomly determines what to do without any policy or principle. A person who makes a rational decision about what he ought to do will reach his conclusion through logical reasoning that follows some guiding principle, or, even if he does not go under the long process of logical inference, he will be led to do some act by following a certain policy. This being the case, if we ask him the reason why he will perform that act, he will be able to provide an answer. Thus, Sidgwick begins by identifying various 'ultimate reasons for actions' which are generally regarded as valid. An *ultimate* reason means a reason which does not call for further reasons. Sidgwick does not explore just any kind of reasons but those which are ultimate reasons for actions because he is searching for the *fundamental* policy to determine what

one ought to do. In addition, Sidgwick confines himself to a discussion of the reasons that are *widely considered* as *valid* (*ME* p. 8). He did so partly because he wanted to concentrate his efforts on a manageable number of methods of ethics by articulating only the methods that are common to us, and partly because he was concerned with the objective rightness of action, namely, the rightness which can be acknowledged not only by the agent himself, but also by other people.

We can explore ultimate reasons by tracing back the reasons for our decisions about what we ought to do (*ME* p. 6 ff.). Some of our 'ought' judgments are hypothetical, implicitly presupposing the targeted end. In such a hypothetical judgment, if a person is asked why he ought to do a certain act, he would answer that he ought to do it in order to accomplish a certain end. Suppose, for example, that an art teacher told his student 'you ought to use this color'. When asked the reason for saying this, the teacher will reply that it is in order for the student to draw a beautiful picture. That one ought to draw a beautiful picture, however, is not an ultimate reason, which people widely recognize as valid. The student may not share the end to draw a beautiful picture, and may not be convinced by the reason just given. In such a case, we have to either admit that this reason is not widely recognized as valid or give a further reason why this student ought to draw a beautiful picture. When we delve into the reasons for actions in this way, we sometimes find an ultimate reason beyond which we cannot go. For example, most of us cannot be indifferent to a person who intentionally decreases his own happiness. We normally think that one ought to look out for one's own happiness, and in our mind this emerges as a categorical prescription which does not necessitate further justification. When we are asked why one ought to ensure one's own happiness, many of us would just say that it is because we ought to take care of ourselves, and we cannot give further reasons. Then, this final statement that one ought to care about one's own happiness is said to be one of the ultimate reasons for actions that are commonly considered as valid.

The selfish concern is not the only reason that we think is ultimate and valid, however. Many people tend to think of such duties or virtues as truthfulness and faithfulness (in Sidgwick's terms, veracity or good faith) as binding without qualifications, regardless of their ulterior consequences (*ME* p. 7). When asked the reason why they ought to keep promises or tell the truth, ordinary people usually answer that it is because they should not break a promise or tell a lie and that is all there is to it. To generalize, they should do it just because it is their duty. Most people recognize such a reason as valid and ultimate.

However, there may be some other reasons for actions that a few people will recognize as ultimate and valid. For example, some people may sacrifice their health, wealth, or even their happiness for the sake of their honor and when asked why, they may answer that they do so just for honor's sake. Sidgwick claims that he will not discuss such a reason because it is not widely recognized as an ultimate reason for what one ought to do. A person *may* seek honorability without further reasons, but few people claim that one ought to act purely for honor's sake. If the claim that one ought to act honorably sounds truly valid and convincing, it is only when there are further reasons for doing so, that is, when an honorable act can generate one's own or other people's happiness, or when it will demonstrate the agent's personal excellence.

Thus, when we carefully examine the reasons for our actions, we can give the following four types of answers as the ultimate reasons that are widely supported as valid (see *ME* Bk. 1 Ch. 1, especially Sec. 4 p. 9):

1. in order to attain moral or intellectual excellence or perfection of human nature,
2. in order to achieve one's own happiness,
3. in order to achieve people's happiness,

or

4. because it is a duty which is prescribed unconditionally.

Here I need to explain Sidgwick's meaning of excellence or perfection, which I have not mentioned in the previous paragraphs. By 'perfection or excellence of human nature' he means attaining an ideal or nearly ideal set of mental qualities, which we admire and approve when they are manifested in human life (*ME* p. 10 fn. 4). According to Sidgwick, virtue is usually regarded as the most valuable element of excellence (*ME* p. 11). Therefore, becoming a virtuous person (a morally ideal person) is included in the ideal of excellence.

Now, if a person acts for the ultimate end of achieving one's own happiness (2), we will consider him as utilizing a method of egoism. An action done for the ultimate end of achieving people's happiness (3) can be regarded as an example of the method of utilitarianism. An action done for duty (4) is based on the intuitionist view, and its systematic method is regarded as that of dogmatic intuitionism, because an action performed out of a sense of duty is usually following certain general rules. When a person acts in order to pursue excellence or perfection in human nature (1), his action exemplifies the method of dogmatic intuitionism.[2] This is because the most important element of excellence

or perfection is virtue, which is manifested by actions that follow general rules of duty and virtue. Hence, the three methods of ethics, which determine what one ought to do, are identified by investigating the four ultimate reasons for actions.

Sidgwick also examines other possible reasons for actions, such as 'the conformity to God's will', 'self-realization', and 'life according to nature' (*ME* Bk. 1 Ch. 6 p. 79 ff.), but he finally rejects those reasons as not ultimate and valid ones. First, as to God's will, there is the practical question of how to correctly determine the true divine will. On the one hand, if God's will is revealed by supernatural intervention, it goes beyond the scope of our study. On the other hand, if our reason can ascertain the divine will, we only need to examine what our reason tells us to do. In addition, it is often said that God's will is human happiness, perfection, or the conformity to duty or virtue. If this is true, our discussion of God's will turns out to be covered by the exploration of the four ultimate reasons already described. If God wills other things than human happiness or perfection, it would be either the self-realization of a human being or a life lived according to nature, the latter two being common ideas of God's will in Sidgwick's time. But Sidgwick denies that self-realization is the ultimate reason for human actions (*ME* Bk. 1 Ch. 7). In the first place, the meaning of 'self-realization' differs from person to person, and what state of affairs it denotes seems ambiguous. It may be said that we realize ourselves 'by exercising, each in its due place and proper degree, all the different faculties, capacities, and propensities, of which our nature is made' (*ME* p. 91), but this does not tell us what 'due place' or 'proper degree' means. Some may believe that a person should choose proper actions by considering his inborn disposition, but no one would recommend he maintain or develop his innate tendency if it is likely to bring about unhappiness to himself or to others. If this is true, we actually regard happiness rather than the exercise of one's natural propensities as the ultimate reason for actions. If, on the other hand, we say that we should exercise those faculties or abilities in accordance with some ideal, it is substantially the same as acting for the purpose of one's perfection or excellence.

The meaning of 'to act according to nature' is also quite ambiguous. All the impulses that we have when we are acting are natural in a sense, but we do not usually think that we may act according to any impulse. If we really believe there is no problem yielding to an impulse, we will not have to worry about deciding what we ought to do. Therefore we need to define a much narrower sense of 'our nature' or 'natural impulses'. Generally, those who claim that we ought to act according to our nature

seem to mean by 'one's nature' those human abilities and disposition which one cannot easily discard, such as (1) propensities which are normal and not extraordinary, (2) those which are innate, or (3) those which are implanted as a result of one's physical makeup or social conditioning. There is no reason, however, why one should not act according to extraordinary or acquired action-tendencies. For example, a love of knowledge or an act of philanthropy is commonly admired despite the fact that they are based on rare and acquired impulses. Moreover, however carefully we observe our physical makeup and our resulting tendencies to action, we cannot answer the question of whether we ought to act according to such tendencies. Likewise, however intensely we observe our social conditioning and its influence on our internal dispositions, we cannot answer the question of whether we ought to follow such dispositions. We do not consider it a fundamental moral principle just to obey social customs. When we think we ought to obey them, we usually take into account social happiness or human perfection as a reason for abiding by such customs.

As another interpretation of 'a life according to nature', some of Sidgwick's contemporaries made a popular claim that we ought to make full use of our potential to realize an ideal society. They asserted that we can do so by forecasting the future of human evolution and making efforts to advance toward the final stage of our evolution. Under the influence of what was called social Darwinism, this type of claim was quite popular worldwide in Sidgwick's time. Sidgwick points out, however, that this claim is groundless, confusing 'what we will be' with 'what we ought to be'.

Thus, Sidgwick rejects these alleged ultimate reasons for action, other than the four reasons listed above. There may still be other reasons for action to be considered, such as 'to act in order to realize freedom', 'to act so that one can adjust oneself to others', etc. However, they are not ultimate reasons if they presuppose further reasons such as 'we should realize freedom to bring about happiness' or 'a person ought to be on good terms with others so that he can concentrate on his own work and reach his excellence or perfection'. If it is alleged that these reasons suggest unconditionally binding rules of conduct, for example, if one considers that one's absolute duty is to realize freedom or to adjust to others, they are substantially the same as the fourth reason shown above. Therefore, the ultimate reasons for action can be classified into any of the four reasons already described, which end up as one of the three methods of ethics, the methods of egoism, utilitarianism, and dogmatic intuitionism.[3] Since we are mainly concerned with the practical procedure of deciding what

we ought to do, what Sidgwick examines is not the four reasons for action but the three methods to determine our actions.

The propositions that state the ultimate reasons for action, on which the three methods of ethics are based, are called the *principles* of ethics.[4] The propositions that 'one ought to aim at one's own happiness', 'one ought to promote people's happiness', and 'one ought to abide by the rules of duty or virtue' are the simple expressions of the principles of egoism, utilitarianism, and dogmatic intuitionism, respectively. When one holds an egoistic view, one adopts the principle of egoism, and decides one's actions using the method of egoism. The same is true of utilitarianism and dogmatic intuitionism.

Some may feel that it is strange to include egoism into the methods of ethics, but as I explained before (1.1 of the present book) Sidgwick takes ethics in a broad sense, and egoism is counted as a category of ethics as far as egoism tells us what we ought to do. Sidgwick intentionally defines ethics in this broad sense because he believes that, however we explain the relationship between egoism and ethics, we always have to face the question of whether to be egoistic or to be morally right. Even if we exclude egoism from the definition of ethics, we cannot eradicate this crucial question.

Some may still think that there could be other methods of ethics. We should note, however, that Sidgwick says that the methods of ethics are only 'conveniently' classed under three heads (see the heading of Book I, Chapter 6, Section 3 of the contents of *ME*). This classification may not be perfect in that there could be still other types of ethical reasoning which use quite different but tenable logic. Nevertheless, for the purpose of Sidgwick's argument, it would suffice to classify our ethical reasoning using those three methods because they adequately reflect problems that concerned most people of his time. The main disputes among moralists of his time were between utilitarianism and intuitionism; the conflict which vexed ordinary people was whether to act for their own happiness, for people's happiness, or out of a sense of duty. In view of the fact that the biggest controversy today is also over consequentialism and nonconsequentialism (deontology and virtue theory), Sidgwick's identification of three methods should still be regarded as valid and significant because they address the central ethical questions of all times.[5]

As I repeatedly point out, the three methods described above, egoism, utilitarianism, and dogmatic intuitionism, are three distinct patterns of reasoning that one and the same person uses in his daily moral thinking, rather than three separate positions each of which

has its own supporters. In reality, there are very few people who are completely egoistic or completely utilitarian. Many people usually use a combination of these three methods, being unable to clearly discern among them. Still, the above three methods abstracted from our common moral thinking are apparently mutually distinct, and often seem to conflict with each other. 'A conflict between two methods' means a situation in which the two methods prescribe different courses of action when applied to a particular case, so that two or more courses of action seem to be right at the same time. In such a conflict, one cannot easily decide what one really ought to do. For example, suppose that a person working for a company detected serious wrongdoing and started to suffer from the decision between whistle-blowing and protecting himself. He may think that he ought to disclose the company's confidential information for the sake of people's happiness, which will be damaged by the company's activities. At the same time, he may also think that he ought to remain silent for the sake of his own lifelong happiness because the whistle-blowing would risk his current and future position. Here, we can say that the method of utilitarianism and that of egoism are in conflict with each other. Note that the principles on which the two methods are based do not necessarily directly contradict each other. As I will explain later, the principle that one ought to promote one's happiness and the principle that one ought to promote people's happiness are mutually independent and do not directly contradict each other. However, even if two principles are not in head-on conflict, two methods derived from those principles could prescribe incompatible courses of action, and such an incompatibility is called a moral or practical conflict.[6]

At this point, we have just identified the three methods which people usually regard as ultimate and valid, and have not yet determined whether they are truly valid; nor have we decided which method overrides the other when two methods come into conflict. In addition, Sidgwick presented these three methods as seemingly distinct but it is no wonder if, on close examination, we find that one method is actually theoretically dependent on another method. Actually, it will later turn out that we cannot fully systematize the method of dogmatic intuitionism without referring to utilitarian thinking.

3.2 The use of intuition

It would be relevant to mention here that Sidgwick was a self-appointed 'intuitive' utilitarian, claiming that he supported utilitarianism on an

intuitive basis. His position is, however, quite different from that of moralists called 'intuitionists' in his time. Therefore, we must clarify the different meanings of 'intuition' used in *The Methods of Ethics*.

3.2.1 The narrower and the wider senses of 'intuition'

The term 'intuition' used by typical intuitionists in Sidgwick's time meant a human faculty which apprehends the rightness or wrongness of an action by looking directly at the action itself, apart from the evaluation of its ulterior consequences. In this sense, one's intuition immediately apprehends a certain precept or rule that designates a specific action that one ought to take. Such precepts and rules are perceived to be unconditional dictates that are self-evident and valid.

However, this is intuition in a narrow sense. According to Sidgwick, the term 'intuition', taken in its much wider sense, means apprehending a certain apparent truth immediately, that is, without any inference. This means that a person grasps a certain proposition as self-evident truth that does not need any proof. He grasps it immediately, and not by any induction from experiences or as a result of any other inferences. This wider sense of intuition means that one could apprehend not only the actions that ought to be performed, but also the ends that ought to be aimed at. Therefore, if we perceive the ultimate end of egoism or utilitarianism (i.e., that one ought to pursue one's own or people's pleasure) to be a self-evident truth, we are using our intuition in this wider sense in talking about these two kinds of hedonism. Intuition in a narrow sense apprehends actions to be done while never considering their ulterior consequences; whereas intuition in a broad sense can also apprehend a self-evident claim about the ultimate ends, that is, the ulterior consequences that our actions would bring about (*ME* Bk. 1 Ch. 8 Sec. 1). In addition to the actions to be done and the ends to be pursued, intuition in this wider sense can also apprehend some abstract axiom that would make up the basis of moral thinking. Actually, Sidgwick posits a theoretical foundation of utilitarianism by intuitively apprehending certain abstract axioms, and this is the reason why he calls himself an intuitionist in this wider sense.

However, it should be pointed out that, whether in a narrow sense or in a wide sense, when he says a certain apparent truth is 'intuitively known', Sidgwick does not presuppose its ultimate validity. It just means that a *seemingly* true precept, rule, end, or axiom is immediately known *as if* it is self-evident and without any inference. Sidgwick admits the possibility that such an intuition may contain flaws and need to be corrected or discarded (see *ME* p. 211).

3.2.2 Induction and intuition

According to the explanation in the previous section, apprehending truth immediately through intuition contrasts with induction by which truth is abstracted from experiences. Nevertheless, even hedonism, which is generally classified under empiricism or inductivism, is a kind of intuitionism in a wider sense. The reason why this holds without contradiction is because, in hedonism, what is known by intuition and what is known by induction are different. In hedonism, a person forms an expectation that a certain action will bring about pleasure through induction based on experiences. One cannot know, however, from induction the principle that pleasure is the sole rational ultimate end of human actions. If someone claims that this principle is true and valid, it must be because he immediately knows that it is true, or because he derives its truth from an assumption intuitively known to be true and valid. In short, in hedonism, one intuitively apprehends the apparent truth that pleasure is the ultimate end, and then ascertains through induction which action is most likely to bring about pleasure. Therefore, hedonism is opposed to intuitionism in the narrow sense, which intuitively determines the right action regardless of its consequences, but not opposed to intuitionism in a wider sense.

In yet another sense, intuition and induction do not oppose each other. According to Sidgwick, we could use an inductive method to explore an intuitive truth (*ME* Bk. 1 Ch. 8 Sec. 2). The apparent validity of the rules of duty, fundamental principles, or ultimate ends in ethics may not be clearly recognized in our daily lives, but may be intuitively apprehended once those principles or ends are precisely formulated. Therefore, we may formulate what can be considered as promising candidates for valid rules, true fundamental principles, or ultimate ends utilizing induction from our own experience or from the commonly held opinions of others before we ascertain their (apparent) validity using our intuition. In such a case, again, there is no contradiction because what can be obtained by induction (several candidates for intuitive truth) and what can be obtained by intuition (which candidate is intuitively valid) are different. In fact, we have already used this inductive method when we identified the three methods of ethics. That is, we first reflect on our experiences to identify what we usually regard as ultimate reasons for action, and then, we decide which ones are truly ultimate and valid by using our intuition. Actually, Sidgwick quite consistently uses this inductive–intuitive logic throughout *The Methods of Ethics*. Most of his arguments utilize the same strategy, in which he closely examines the data that are obtained from our experiences, then

gradually weeds out inadequate candidates in the light of his and other people's intuition, and finally reaches intuitive truth that seems truly self-evident and valid.

3.2.3 Three phases of intuitionism

Intuitionism in the wide sense, in which one intuitively apprehends every kind of apparent truth, can be classified into three phases according to the types of apprehended truth. They are termed perceptional, dogmatic, and philosophical intuitionism (*ME* Bk. 1 Ch. 8).

Perceptional intuitionism corresponds to the previously mentioned view that tells one to act according to one's conscience. In this phase of intuitionism, one's intuition is supposed to apprehend the rightness or wrongness of each act in a particular situation. In dogmatic intuitionism, it apprehends certain general rules by which one can judge the rightness or wrongness of actions. In philosophical intuitionism, it apprehends much more abstract axioms which form the foundations of our ethical reasoning. These three cases are all called intuitionism in its broader sense, because in every phase one immediately apprehends a certain (*apparent*) truth. However, perceptional and dogmatic intuitionism only use intuition in a narrow sense, both emphasizing that one can determine the right action without regard to the action's ulterior consequences. All three phases can be found in the moral reasoning of ordinary people. Most of the so-called intuitionists in Sidgwick's time claimed either perceptional or dogmatic utilitarianism, whereas Sidgwick himself adopted philosophical intuitionism.

Of the three phases of intuitionism, it is only dogmatic intuitionism that Sidgwick examines as one of the methods of ethics. This is because, first, philosophical intuitionism does not present practical procedures for determining actions in particular cases, but offers more philosophical, fundamental axioms which guide our moral reasoning. Second, perceptional intuitionism is unsuitable for a systematic study of the methods of ethics for the reasons stated, even though it admittedly exhibits the process by which one decides the rightness or wrongness of an action. According to perceptional intuitionism, we only need intuition to decide each particular action, and we will be dissatisfied with this logic as long as we seek a clear, consistent, and convincing guide to determine what we really ought to do. We cannot always ascertain what our conscience orders us to do. What conscience tells us to do can differ from person to person. Even in one person's mind, what his conscience tells him to do can change over time. Thus, we turn to ethics because we cannot always trust our conscience. It is dogmatic intuitionism rather

than perceptional intuitionism that may presumably provide us with a systematic and coherent guide for actions. Therefore, the method of dogmatic intuitionism is identified as one of the three methods of ethics that Sidgwick examines in *The Methods of Ethics*.

Sidgwick, however, later realizes that the method of dogmatic intuitionism cannot actually provide a very clear guide for actions after all, and moves on to the philosophical phase of intuitionism in order to explore more abstract fundamental principles. As a result of such an investigation, Sidgwick claims that by using sophisticated philosophical intuition we can find the foundations of utilitarianism. This is how he came to call himself a utilitarian on an intuitive basis.

3.3 'Commonsense' and the 'morality of commonsense'

There is another important device that Sidgwick uses when he identifies three methods of ethics. This is what he calls commonsense. 'Commonsense' is an important notion in Sidgwick's theory because it is the starting point of his speculation and provides additional support for the validity of his arguments.

Commonsense generally means the feelings, understandings, judgments, or opinions that all or most of us humans share (Sidgwick frequently uses the phrase 'commonsense of mankind'). However, Sidgwick seems to use this term with two different connotations, though he himself does not explicitly state such different uses of this term.

First, it means the widely accepted opinions, which we *already* have in common. Sidgwick extracts the three methods of ethics from our 'commonsense', and tries to examine 'common' moral rules from this point of view. The 'Morality of Common Sense', or 'common morality', is a set of general rules of conduct, which is thought to embody moral truths, and is recognized as such by the consensus of mankind – or at least of the people who have adequate intellect and serious concerns for morality (*ME* Bk. 3 Ch. 1 p. 215). A set of rules that is publicly recognized as what individuals ought to observe is called 'positive morality', just as a set of existing laws, which have been publicly posited by people, is called positive laws. Among such rules of positive morality, the ones that all or most people would agree to endorse as truly valid are the rules of common-sense morality. People are supposed to share such a set of rules, and to appeal to it in their moral discussions (*ME* p. 216). In short, commonsense morality means the whole body of moral rules that are commonly used and recognized as sound by people in general. Such rules of common morality are usually put in terms of particular duties and virtues.

According to dogmatic intuitionism, one regards the morality of commonsense as valid in and of itself, by intuitively judging that one ought to unconditionally observe such duties and virtues (on the assumption that we can clearly formulate each duty and virtue). It is evident that the term 'commonsense' in this context is the first meaning of commonsense. However, it is possible that existing moral rules, even though we do share and usually consider them as valid, may turn out not to be valid *in themselves*. Actually, Sidgwick later criticizes the method of dogmatic intuitionism and concludes that common-sense morality cannot be valid apart from situations and various other considerations.

We can say, however, that our conclusion as to the invalidity of common-sense morality is actually based on the common feelings, senses or opinions shared by those who are reflecting on it. This forms the second meaning of commonsense. In other words, 'commonsense' sometimes means the opinions or feelings which people *would* commonly hold *after reflection*. Sidgwick sometimes calls it 'Reflective Common Sense' (see, for example, Bk. 3 Ch. 8 Sec. 1 of the contents of *ME*7). Reflective commonsense appears as (philosophical) intuition that a person would have if he undergoes reflection, and as opinions that reflective people would share. Sidgwick gives special importance to this reflective commonsense, and often utilizes it in order to ascertain the soundness of his own argument, and to persuade people. At the same time, however, our reflective commonsense is not infallible. It is logically possible that our intuition or opinions may still contain errors that we cannot perceive even after a long process of reflection. Therefore, the mere fact that a certain argument perfectly suits our reflective commonsense does not completely assure the validity of that argument. Nevertheless, if an argument severely goes against our reflective commonsense, this may well constitute a warning against the plausibility of that argument. In contrast, if we ascertain that an argument does not conclusively oppose our reflective commonsense, it may well offer a partial support for that argument (even though it does not perfectly prove the truth of that argument), in the sense that we have not discovered the grounds to reject that argument.

Using these two interpretations of 'commonsense' as crucial touchstones of his arguments, Sidgwick makes his inquiry into ethics. We can think of several reasons why he took this line of argument.

First, Sidgwick's purpose in *ME* is to systematize our daily moral reasoning. Naturally, the starting point of his inquiry is the different patterns of moral thinking that can be found in our commonsense (in the first meaning of this term), and the various decision-making procedures abstracted from them.

Second, common-sense morality is what most people admit to have general validity and binding power. Philosophers and moralists are no exception here; even people of the intuitionist and utilitarian schools in Sidgwick's time, whose theoretical stances severely conflicted with each other, accepted the rules of common-sense morality as more or less valid and binding. They were also willing to reexamine the validity of their own theoretical positions in the light of common-sense morality. Therefore, it sounds reasonable for us to first examine common-sense morality in our search for truly valid ethical theory, and to refer to it whenever we reflect on our moral thinking. Besides, if we can establish a theory that neatly systematizes our common-sense morality while clarifying its philosophical foundations, a wide range of people will be very likely to accept that theory.

Third, reflective commonsense can also be used to check the soundness of an individual's intuition. Actually, Sidgwick supports his argument regarding the deepest foundations of ethics by using his own philosophical intuition, but he is also aware that his philosophical intuition can turn out to be false or incorrect, as previously suggested. Sidgwick believes that the only way to ascertain the soundness of one's own intuition is to refer to our reflective commonsense. Of course, our reflective commonsense itself can turn out to be false, so it cannot perfectly prove the legitimacy of his argument. Still, we can assume that an appeal to our reflective commonsense would make it easier for a wide range of people to accept his argument, because reflective commonsense is a feeling that most people would share when they undertake reflection. (Similar points are suggested by several other writers, though my explanation is a little different from theirs. See, for example, Schneewind 1977, pp. 191–3, 262; P. Singer 1974; Shionoya 1984, p. 149.)

In any case, Sidgwick appeals to our commonsense because he wishes to systematize the morality which we now hold in common, and to make it more persuasive for ordinary people. Thus, in *The Methods of Ethics*, (1) the three methods of ethics are abstracted from our commonsense (i.e., the ways of ethical reasoning we already have), and (2) the method of dogmatic intuitionism is criticized in his examination of common-sense morality. Once Sidgwick apprehends the fundamental principles of ethics using his own philosophical intuition, (3) he examines the validity of these principles in the light of our reflective commonsense. The apparent soundness of other philosophical analyses and arguments is also ascertained in the same way. Finally, through such philosophical arguments (4) our common-sense morality is reconsidered and systematized.

4
Meta-Ethical Analyses

According to Sidgwick, ethical judgments are, primarily, judgments about the *right* actions, or actions which *ought* to be done. Therefore, it is essential for students of ethics to understand the meanings of the terms 'right' and 'ought'. At the same time, the notion of 'good' also plays an important role in ethics. Moral actions are often called 'good' acts. Some actions are judged as right because they attain a certain ultimate end, which is also called a human's 'True Good' (*ME* p. 3). Moreover, two of the three axioms, which Sidgwick proposes as the fundamental axioms to determine the right actions, tell us to aim at certain kinds of good as the ultimate ends of human actions. Therefore, the basic moral concepts that we need to analyze are those of 'right', 'ought', and 'good'.

We should also note that Sidgwick considers the right actions, or the actions which ought to be done, as 'rational' actions. Sidgwick says that the ultimate ends are ordered by our reason. The adjective 'rational' is also used in his names for two fundamental principles, which are the Principles of Rational Self-Love and Rational Benevolence. Therefore, we had better understand Sidgwick's meaning of '(the faculty of) reason' and 'rational'. I will first explain the concept of reason and rationality, and then go into analyses of the concepts of right and good.

4.1 Sidgwick's meaning of 'reason'

There is no clear definition of the faculty of reason, or rationality, in the seventh edition of *The Methods of Ethics*. However, it is possible to understand what they mean by analyzing the contexts in which these terms are used.

What Sidgwick means by the term 'reason' in the context of ethical thinking, or 'Practical Reason', is the faculty to recognize a certain truth

about what is right or wrong. The cognition of such a truth urges one's will, by providing motives or impulses toward actions, or impulses to aim at certain objects. Such an urge is called 'the dictates of reason' (see, for example, *ME* Bk. 1 Ch. 3 pp. 34–7). Truth is what cannot turn to falsehood without a reason. Likewise, a proposition which contradicts 'truth' cannot turn out to be true for any reason (*ME*7 p. 34; *ME*1 p. 23). Most people think 'truth' can be universalized, in that it applies not only to one particular case but also to all similar cases. It is also considered to be 'objective' in the sense that it does not change according to a single person's whimsical feelings and moods and that ideally intelligent and reflective persons would similarly admit its veracity (*ME* pp. 341, 399). To 'recognize' truth means that an individual apprehends such an objective truth immediately, or intuitively in the broad sense. Here, of course, nobody knows whether such an apprehension is really valid, as this cognition may turn out to be fake or mistaken. Sidgwick points out that, to be exact, this cognition should be called *apparent* cognition (*ME* p. 34 fn. 2). This implies that cognition is not always valid, but at least it is certain that such an intuitive apprehension occurs as a matter of psychological fact, and that the person who apprehends it tends to claim that his apprehension is true and valid. This cognitive faculty recognizes something as truth, which cannot turn into falsehood unless some error or overlooked factor is discovered. Therefore Sidgwick calls this ability 'reason' rather than 'a moral sense', because the latter could mean an ability to have a certain feeling that can change, depending on the agent's state of mind and inclinations.

According to Sidgwick, the term 'reason' is used in several different contexts. First, reason is considered as (1) a faculty of logical reasoning – in Sidgwick's words, 'conversant in its discursive operation with the relation of judgments or propositions' (*ME* p. 34 fn. 1). But, second, reason is often said to be (2) a faculty to apprehend universal truths intuitively – that is, apprehend them as self-evident and without reference to any further premise. Reason in this second sense is called 'intuitive reason' (*ibid*. Sidgwick suggests that, even in the field of logic, the axioms of logic and mathematics are apprehended by intuitive reason). Sidgwick believes that, in ethics, the latter, intuitive reason is even more important than the former. This intuitive reason apprehends (2-i) the self-evident truth that 'it is right to adopt appropriate means to an end' and (2-ii) duties that one ought to do, or ends that one ought to target. These intuitions give us orders such as 'Adopt this means', 'Do this duty', 'Aim for this end', etc. In short, the faculty of reason has three functions: one, logical reasoning, two, ordering appropriate means to a

given end, and three, ordering duties or ends in themselves. Of course, if reason orders an end, reason goes on to order an appropriate means to that end.

Such a concept of reason may seem to contrast with the Humean view of rationality, which was well known among British philosophers in Sidgwick's time. The function of reason that Sidgwick has in mind is not restricted to one of choosing appropriate means to an end as Hume claimed. Moreover, Sidgwick's claim that reason gives motives for actions is also contrary to Hume's view, which interprets reason strictly as the faculty in charge of judgments about truth and falsehood and claims that reason never becomes a motive for action.

Sidgwick's analysis of rationality starts with examining how this concept is held in the common view and locution of ordinary people, and by different schools of moral philosophy.

What ordinary people immediately associate with the term 'rational' is an experience commonly called 'a conflict between reason and non-rational impulses'. When a person is driven to act against his well-considered judgment, he is said to have an irrational desire or motive.

> Every one, I suppose, has had experience of what is meant by the conflict of non-rational or irrational desires with reason; most of us (e.g.) occasionally feel bodily appetite prompting us to indulgences which we judge to be imprudent, and anger prompting us to acts which we disapprove as unjust or unkind. It is when this conflict occurs that the desires are said to be irrational, as impelling us to volitions opposed to our deliberate judgments.
>
> (*ME* Bk. 1 Ch. 3 Sec. 1 pp. 23–4. A similar point is made in Sidgwick 1893b.)

We feel the impulsive force of such irrational desires most clearly when we do not yield to them but rather resist them based on deliberate judgments, because in such cases we need to make voluntary efforts to overcome those irrational desires. When we act according to our impulses without passing any deliberate judgment, our impulses and actions would be more appropriately called nonrational.

These conflicts between reason and nonrational or irrational desires are conflicts between deliberate judgments, which have attained cognition through the deliberation process, and desires that go against them. We usually think that we can show, by argument, that a certain action is irrational, and this suggests that we believe there is something which all of us can recognize as true and valid, and which we can use

to persuade other people. Such cognition can properly be called cognition of apparent truth. To put it more precisely, however, these conflicts should be regarded as competitions between those *motives* or *impulses* aroused from cognition-based judgments and those *desires* that go against them. We feel these conflicts because our deliberate judgments are accompanied by powerful motives which pull us in the opposite direction from desire which drives us to perform certain irrational acts. Additionally, we believe that we can convince others, by appealing to their reason, to perform different acts than those they intended – even though they may not actually follow our suggestions. Sometimes we try to persuade people to change their moral opinions, by appealing to their reason. This suggests, according to Sidgwick, that reason does have motivational power, in conjunction with its function of recognizing apparent truth.

Now, what kind of cognition tends to motivate people to refrain from performing nonrational or irrational actions?

Some so-called nonrational or irrational actions also involve a certain cognitive or intellectual process. Certainly, there are instinctive or reckless impulses in which one is unconscious of the means to attain an end or even of the end itself; but, as far as we are concerned with the 'voluntary actions' discussed in ethics, the agent is usually aware of the expected results (or ends) that he seeks, and the means to attain that end before he determines the will to perform the act in question. Therefore, reason can utilize this intellectual process in two ways, in order to deter irrational actions. First, it can help a person see the proper means to attain the end he desires. Second, it can help a person recognize a certain present or expected fact that will arouse a new desire or aversion in him – for example, by making him vividly realize the expected (good or bad) consequences of the action he is attempting to do.

Sidgwick believes, however, that the function of reason is not limited to that of presenting appropriate means or relevant facts. If reason does only these things, our desires may or may not be revised by those presented facts. As a result, we may only act according to the strongest impulses at the time of action. However, we actually believe that it is *irrational* not to adopt indispensable means to a given end. In believing this, we are not just recognizing the *fact* that 'this means is indispensable for attaining the end', but getting an *imperative* to adopt an indispensable means to an end. This *motive* to adopt appropriate means to a given end, or the *will* urged by this motive, is a higher motive or will, distinct from a desire or a will to do a particular act in a particular situation. This higher motive is formed by apprehending the apparently

self-evident truth that it is right to adopt indispensable means to a given end. We also have other types of higher wills, such as a resolution to act in a certain way at a certain future point, or a will to adopt a certain end as supreme, and they also seem distinct from evanescent motives to perform particular acts. Such higher wills are more constant ones, and we form these wills by recognizing certain apparent truths that are independent of particular situations. Thus, according to Sidgwick, practical reason is the faculty to generate enduring motives based on the recognition of certain truths, which urge us to determine higher wills which are distinct from momentary motives or impulses.

We feel such higher motives based on the recognition of truth as quite different from other transient impulses or desires. In an individual's mind, this special motive emerges as a 'dictate' of reason, which has a prescriptive power to suppress other impulses. However, we cannot always control our acts by such rational motives. The motive our reason gives us is but one among many motives, and is not always dominant. This is why the agent often falls into conflict between a rational motive and other desires.

> According to my observation of consciousness, the adoption of an end as paramount [. . .] is quite a distinct psychical phenomenon from desire: it is a kind of volition, though it is, of course specifically different from a volition initiating a particular immediate action. As a species intermediate between the two, we may place resolutions to act in a certain way at some future time: we continually make such resolutions, and sometimes when the time comes for carrying them out, we do in fact otherwise under the influence of passion or mere habit, without consciously cancelling [*sic*] our previous resolve. This inconsistency of will our practical reason condemns as irrational, even apart from any judgment of approbation or disapprobation on either volition considered by itself. There is a similar inconsistency between the adoption of an end and a general refusal to take whatever means we may see to be indispensable to its attainment [. . .] such a contradiction as I have described, between a general resolution and a particular volition, is surely a matter of common experience.
>
> (*ME* Bk. 1 Ch. 3 Sec. 4 pp. 37–8)

Interestingly, Sidgwick delineates the notion of 'reason' much more clearly in the first edition of *The Methods of Ethics*, rather than its seventh edition. The title of Book I, Chapter 3 of the first edition is 'Moral Reason', whereas that of the seventh edition is 'Ethical Judgments'.

In the table of contents of the first edition, two sections of Book I Chapter 3 clearly state that reason is the faculty to apprehend truth, and that moral reason is a source of action.

In addition, the main text of the first edition explains that reason gives consistency to actions. It even explains that, ideally, we cannot say that we have given full play to our practical reason unless the complete order, harmony, and unity of a system is introduced into our actions (*ME*1 p. 26; Schneewind 1977, Ch. 7 Appendix. We can also read in the passage cited above a nuance that reason seeks consistency). Apparently, inconsistent actions are irrational. 'Inconsistent' here means that one is driven by two mutually conflicting impulses, or that one has different impulses which are not systematized. So, a person is called irrational in the sense that he has two conflicting impulses (1) when he rejects the appropriate means to the end which he desires. Also, (2) when a person generally aims at a certain kind of object or result and yet does not deliberately aim at a particular object or result which belongs to that category, he is regarded as contradicting himself or having unsystematic impulses, and hence called irrational. In addition, those who claim that they are acting rationally usually act according to general rules or concepts. Such people think of those impulses as irrational (3) that are contradicting those general rules, or (4) that do not come under any basic rule. Moreover, it is irrational to act according to a general rule (5) when it conflicts with another general rule or when it is not systematized. Thus, reason comes to undertake the task of determining the ultimate ends and the true first principle of actions, which will systematize and harmonize those impulses or rules (*ME*1 p. 26).

The reason why the title of Book I, Chapter 3 was changed is probably because the phrase 'Moral Reason' gives the impression that Sidgwick presupposes the existence of a faculty which is clearly recognizable. This may evoke a counterargument that we have no evidence for the existence of such a faculty.[1] In the seventh edition of *ME*, Sidgwick refrains from explaining reason as an independent faculty, and adopts a different line of argument, which is that the characteristics of the moral judgments we usually make can be explained as if they are dictated by reason. Reason here is regarded as a sort of expedient term which expresses the properties of moral judgments or judgments of truth and falsehood. This is probably why the term reason frequently appears in quotation marks in the seventh edition. Due to this change, however, the clear description of reason in the first edition disappeared in the seventh edition, which has made it harder for us to understand what Sidgwick means by reason. Nevertheless, it is clear that the characteristics

stated above are those that Sidgwick meant by using the terms rational and reason.

In the following sections, I will argue that ethical judgments are not merely claims that certain feelings (such as moral sentiments) exist, but that ethical judgments are related to the cognition of apparently self-evident truths, and that such cognition gives us certain motives or impulses. When such motives conflict with other impulses, we feel the force to suppress those impulses. The term 'rational' is used to express such unique characteristics of our ethical judgments.

4.2 'Ought' and 'right'

The terms 'ought' and 'right' have several properties in common, and in most cases they are interchangeable. Both terms are primarily applied to actions. Statements such as 'one ought to do this act' or 'this is the right action' contain authoritative prescriptions that tell us to do a certain act, either unconditionally or for certain purposes.

Sidgwick explains the characteristics common to these concepts as follows. First, the concepts expressed by 'ought' and 'right' are totally different from factual concepts. Second, these concepts are simple and indefinable. Third, they express rational judgments based on some cognition of apparent truth. This recognition of truth or reasonableness supplies the motive for action. Thus, we often succeed in convincing people to do what they ought to do by argument, by making them aware of some overlooked truth, and thereby helping them acquire a motive to do it.

> [W]e commonly think that wrong conduct is essentially irrational, and can be shown to be so by argument; and though we do not conceive that it is by reason alone that men are influenced to act rightly, we still hold that appeals to the reason are an essential part of all moral persuasion.
>
> (*ME* Bk. 1 Ch. 3 Sec. 1 p. 23)

To explain more precisely, the judgments in which the term 'ought' or 'duty' is used and those in which 'right' is used are slightly different. The former suggest the existence of desires which go against those judgments, but the latter do not necessarily contain such opposing desires. Sidgwick thinks, however, that what is important is not the difference but the common characteristics described above.

Sidgwick is particularly concerned with the ethical use of the terms 'right' and 'ought'. By ethical use he means those uses which express

'true' duties regarding the actions of an individual.[2] We sometimes say 'we ought to do this because it is required by law', but this is primarily the legal use of 'ought' and not the ethical use of the term – though this could turn out to be an ethical use if one judges that it is one's *true duty* to obey the law. The ethical use of these terms is when we say that we morally ought to do something even if we have no legal obligation to do it. Again, the terms 'right' and 'ought' are sometimes used in sentences in which a nation or a public institution is the subject word, such as when we make a political judgment that our nation ought to adopt a particular policy; but this is not the ethical use of the term in that it is not a judgment about an individual's duty to perform an action (cf. *ME* p. 29; p. 34 fn. 4).

4.2.1 Difference from factual judgments

The concepts expressed by terms such as 'ought' and 'right' are involved in practical judgments. The first point Sidgwick makes is that these practical judgments are *not* the judgments that describe present or future facts. According to Sidgwick, all attempts to explain practical judgments or propositions regarding 'ought' and 'right' without recognizing their unique character are inadequate, 'the fundamental notion represented by the word "ought" or "right" [. . .] being essentially different from all notions representing facts of physical or psychical experience' (*ME* p. 25). In particular, moral judgments, in the sense in which they are distinguished from prudential judgments, are usually accompanied by what are called moral sentiments, and these sentiments often exert certain influences on a person's determination of a will. Some people claim that our moral judgments are merely the expression of such moral sentiments. Yet Sidgwick denies this interpretation, claiming that the fact that one has certain moral sentiments is not equivalent to the statement that one ought to do something. This fundamental difference between mere facts and practical judgments is widely accepted as the fundamental gap between 'is' and 'ought', and I will proceed with this as our basic assumption.

4.2.2 Simple and indefinable

Then, precisely what do 'ought' and 'right' mean? Sidgwick examines several common interpretations.

According to the first interpretation, the statement that an action is right or ought to be done implies that the action in question is the best, or simply an appropriate, means to a given end. This is a cogent interpretation, as expressing the common usage of these terms. By saying

that a certain action is right or ought to be done, we often mean that it is favorable from the viewpoint of the agent's self-interest or happiness. In such cases our judgment actually suggests that a certain action is appropriate as a means to the end of pursuing self-interest. However, we should also note two kinds of exceptions to such a usage. First, a certain kind of action, such as the act of justice or veracity, is commonly judged to be unconditionally right, regardless of its consequences. Second, we also judge that the adoption of a certain end, such as the common good or general happiness, is right in itself. These two types of practical judgments, and especially moral judgments distinguished from prudential judgments, do not match this first interpretation in that they are not concerned with a means–end relationship.

Then, what is the interpretation that can also apply to moral judgments distinguished from prudential judgments? The second interpretation, which Sidgwick examines, is that ought- and right-judgments express the existence of special feelings of approbation or disapprobation in the speaker's mind (*ME* p. 26). This is the interpretation that immediately occurs to us when we recall that moral judgments are usually accompanied by human feelings, called moral sentiments. However, as we already discussed, this is an invalid interpretation which does not capture the fundamental differences between practical and factual judgments. The irrelevance of this interpretation becomes even clearer when we consider the following. Suppose that Person A states that 'Person C ought to tell the truth', while at the same time Person B states that 'C ought not to tell the truth'. If moral judgments only express the fact of the speaker's approbation or disapprobation, this case should be regarded as merely presenting two mutually coexistent facts, namely, the fact that A has the feeling to approve C's telling the truth, and the fact that B has the feeling to disapprove C's telling the truth. These two facts just mean that A and B have different feelings, and do not contradict or conflict with each other. However, we usually think that we cannot regard the above two ought-judgments as both true at the same time. When we make a moral judgment in a certain situation, we usually assume that other people would also make the same judgment in the same situation unless there is a mistake or misunderstanding. In Sidgwick's words, 'The peculiar emotion of moral approbation is, in my experience, inseparably bound up with the conviction, implicit or explicit, that the conduct approved is "really" right – i.e. that it cannot, without error, be disapproved by any other mind' (*ME* p. 27).

Then, the third interpretation comes up by slightly revising the second interpretation. This claims that the moral judgments expressed

by the term 'ought', 'right', or similar ones mean that a certain feeling mixed with sympathy exists in the speaker's mind. According to this interpretation, what these judgments express is 'not the mere liking or aversion of an individual for certain kinds of conduct', but the feeling of approbation or disapprobation that is 'complicated by a sympathetic representation of similar likings or aversions felt by other human beings' (*ME* p. 28). Sidgwick admits that sympathy usually accompanies moral sentiments, and that without sympathy it becomes hard for one to sustain one's moral sentiments. However, one's moral sentiment sometimes goes against the conventional feelings of people in a society as well as those of oneself. In such cases, this moral sentiment conflicts with feelings that are obtained by sympathizing with others, just as it conflicts with one's own liking or aversion. Therefore, moral sentiments are essentially different from sympathetic relationships with various people, and we cannot regard moral judgments that accompany such moral sentiments as expressions of those connections.

The fourth interpretation can be grasped by taking a different perspective. This interpretation, which applies only to judgments that contain 'ought' or 'duty', claims that the statement 'a person ought to do this' means that he will be sanctioned by the public unless he performs that act (*ME* pp. 28–30). This interpretation surely contains a part of the meaning of 'ought' or 'duty'. We commonly regard duty as what is expressed by the term 'ought', and in positive law duty is firmly connected to punishment. Sidgwick, however, points out that this interpretation does not necessarily apply to the special ethical use of the term 'ought'. What we mean when we say 'we morally ought to do this even though we have no legal obligation to do so' is not simply that we will be punished unless we do it. When we say 'he ought to be condemned' while we know that he will not actually be punished, this 'ought' cannot mean that someone may be punished.

Finally, as a variation of the fourth interpretation, there is a claim by religious advocates that the ought-judgments mean that one will be punished by God if one does not perform the suggested act. Not all people share this theistic belief, however. In addition, even those who have this belief do not seem to identify the judgment 'he ought to do this' with the judgment 'he will be punished by God unless he does it'. They are convinced that he will be punished by God unless he does it *exactly because* that is what he ought to do. Also, believers often talk about what is called God's justice, and this implies that what God does is always just and right. But it is obvious that 'right' in this context cannot mean that God will be punished unless God does it.

Thus, after rejecting all these interpretations, Sidgwick concludes that the notions common to the terms 'ought', 'right', and those which express similar basic concepts are so fundamental that we cannot give them a formal definition (*ME* p. 13). He does not mean that these concepts are innately preprogrammed in the human mind, but he claims that, whether they are innate or acquired, in our present thought these concepts appear as fundamental, and that we cannot reduce them to simpler concepts. This means that we cannot define the concepts 'ought' or 'right' in such a form that 'A is C which is B'. We may define the concept of a triangle as 'a shape that has three sides' and that of peach cobbler as 'a baked sweet made of peaches and wheat crusts', but we cannot define 'ought' or 'right' in such a way.

One may question the plausibility of Sidgwick's claim that the concept expressed by 'ought' or 'right' is indefinable. Actually, there is no definite argument to prove the validity of this claim in *The Methods of Ethics*. I will later refer to Schneewind's view that Sidgwick could have defined those concepts (see 4.3.2 of this book). At present, it will suffice to say that, as far as we can tell by Sidgwick's failed attempts to define these concepts, it seems quite difficult to define them. According to Sidgwick, the only thing we can do to clarify the meanings of these concepts is to explain as precisely as possible how these concepts are related to other concepts. That is, Sidgwick gives up the idea of directly defining them and he clarifies their logical or semantic properties by examining how and with what implications these concepts are used. This approach bears a striking resemblance to the method of contemporary meta-ethical analysis such as Hare's, in that it focuses on determining the logical properties of the terms by scrutinizing the common use of them rather than defining them.

4.2.3 The properties of 'ought' and 'right'

Sidgwick thus launches his analysis of the properties of the concepts expressed by 'ought' and 'right'. He first distinguishes two kinds of cases, in which the term 'ought' has different connotations. In one case, the ought-statement implies that a person can do what ought to be done, and in another case he may not be able to do it. Sidgwick confines himself to the former kind of cases, because ethics is concerned with what a person is actually going to do, and hence with his choice among possibilities. This former type of usage is said to be the narrowest ethical meaning of 'ought'.

> [I]n the narrowest ethical sense 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. In a wider sense, however, – which cannot conveniently be discarded – I sometimes judge that 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. And this wider sense seems to be that in which the word is normally used in the precepts of Art generally, and in political judgments.
>
> (*ME* p. 33)

The 'wider' sense can be explained with the following examples. One's claim that 'we ought to restore economy within a year' may mean that one thinks of such a situation as ideal and desirable, and may not imply that one can actualize it by an individual's single effort or people's collective action. Again, a technological judgment such as 'we ought to make aircrafts that will never crash' does not necessarily mean that we can do so in reality. Aesthetic judgments such as 'the Matterhorn ought to always be covered with snow under bright sunshine' or 'there ought to be fewer tourists in the temples in Kyoto in the spring when the cherry blossoms are in bloom' seem to mean that such graphic images are beautiful and desirable, and do not imply that an individual can bring about such changes. These types of judgments are not concerned with actions that an individual can bring about with his or her own volition, and hence irrelevant to ethics, which is our main concern.

However, we can point out one characteristic which is common to all ought- or right-judgments.

> In either case, however, I imply 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.
>
> (*ME* p. 33)

Whether one states 'we ought to make aircrafts that will never crash' or 'the Matterhorn ought to always be covered with snow under bright sunshine', the ought-statements have peculiar nuances that are different from those of statements about one's desire, such as 'I want them to make aircrafts that will never crash' or 'I want the Matterhorn to be

covered with snow under sunshine'. The statements containing 'ought' suggest one's conviction rather than one's merely personal and transient sentiments. They imply that one's judgments have certain grounds that other people would agree with. A person who seriously utters the ought-statement has expectations that any other rational being, who seriously judges that matter, would concur with his own judgment. According to the explanation in the previous section of this book, rational beings are those who have the faculty to recognize truth. Truth means that which cannot be judged differently by different people, provided there are no errors or misunderstandings. By making a judgment that can be restated with the term 'ought' or 'right', a person suggests that he is recognizing what cannot be judged differently by different people, and implies that other rational beings must also recognize the same truth and make the same judgment.

An ethical judgment that implies an individual 'can' perform the action in question, such as 'Person P ought to do action A in situation S', also implies that it should be judged similarly by all rational beings unless an error is pointed out. The person who makes this judgment is convinced of its validity, and his conviction reflects the apparent cognition of a certain universal truth, which will not be swayed by his personal feelings. Such judgments would suitably be called 'rational' judgments, or judgments 'passed by reason'. Sidgwick further points out the following in order to justify himself in calling ethical judgments rational.

> [E]ven when a moral judgment relates primarily to some particular action we commonly regard it as applicable to any other action belonging to a certain definable class: so that the moral truth apprehended is implicitly conceived to be intrinsically universal, though particular in our first apprehension of it.
>
> (*ME* p. 34)

> Further, when I speak of the cognition or judgment that 'X ought to be done' – in the stricter ethical sense of the term ought – as a 'dictate' or 'precept' of reason to the persons to whom it relates, I imply that in rational beings as such this cognition gives an impulse or motive to action.
>
> (*ME* p. 34)

In short, an ethical judgment is (1) a judgment in which, even when a particular person makes it about a particular action, he believes others

would make similar judgments that would apply equally to similar actions. This suggests that this ethical judgment is based on the cognition of a certain universal truth, which is not confined to a particular situation. Additionally, when the action in question is feasible, (2) an ethical judgment provides certain motives or impulses based on such cognition. Humans are, however, not always rational and do not always recognize truth. For this reason, for humans, a motive produced by an ethical judgment is but one among many motives which can conflict with each other, and is not the dominant motive. When the motive based on cognition comes into conflict with motives derived from other impulses, that phenomenon is described as a conflict between reason and nonrational impulses. In such conflicts, the rational motive is called a dictate of reason, and it is expressed by terms such as 'ought', 'duty', and 'moral obligation'. These terms do not apply to the judgment of a perfectly rational being who is without conflicts, but we can still use terms such as 'rational' or 'right'. Thus, ethical judgments express rational judgments in which one recognizes apparent truth and obtains certain motives or impulses based on that cognition.

What is particularly important about Sidgwick's analyses of ethical judgments are the features (1) and (2) shown above. This is because these two logical properties of ethical judgments play significant roles in Sidgwick's arguments concerning fundamental ethical principles. The Principle of Justice, which we will examine later, is derived from feature (1) of the ought-judgment. R. M. Hare also analyzes ought-judgments and points out two logical properties equivalent to (1) and (2), calling them *universalizability* and *prescriptivity* respectively. I will discuss Hare's analyses in Chapter 8 in this book, but here three points should be made. First, Hare's universalizability is considered to be a formal property of evaluative judgments in general – that is, a property that a judgment has as far as it takes the form of evaluative judgments – including 'good' judgments as well as 'ought' and 'right'. Sidgwick's Principle of Justice only corresponds to the universalizability of ought-judgments in Hare's meta-ethical analysis, while Sidgwick gives a slightly different analysis of the concept of good. Second, as I will demonstrate later in this book, Sidgwick points out that the term 'ought' is also used in judgments about means–end relationships and in prudential judgments, and claims that all practical ought-judgments including these have those 'rational' properties described above. In contrast, in his *Moral Thinking*, Hare mainly focuses on the claim that the term 'ought' *in moral contexts* has the two properties mentioned above, but does not expound on the fact that the same term can also be used in judgments of means–end

relationships or in prudential judgments. This is probably because for the purpose of his book only the analyses of 'ought' judgments in moral context were needed to construct his main argument. Indeed, Hare also argues that *all* the ought-judgments have universalizability and prescriptivity, and in his other writings he admits that judgments of means–end relationships and prudential judgments can also have universalizability. We should note, however, that Sidgwick does not distinguish between the moral and nonmoral use of the term 'ought' at all; according to him, they have the very same logical properties (1) and (2) explained above. To him, if moral judgments are different from nonmoral ought-judgments, this is not because their logical properties are different, but because moral judgments are based on additional fundamental principles which are intuitively apprehended by the person who is judging. This forms the most crucial point of difference between Sidgwick's and Hare's meta-ethical analyses and their approaches to construct normative ethical theories. Third, Hare's universalizability of ought-judgments has the same basic meaning as Sidgwick's Principle of Justice, but Hare's principle of universalizability of moral judgments asks us to consider the feelings of others, beyond what Sidgwick's Principle of Justice requires of us. As we will see later, Hare's additional requirement arises not from universalizability, which is a pure logical property of ought-judgments, but because Hare tacitly introduces other elements needed for moral reasoning. This will become clear in Chapter 8 of this book.

4.2.4 Goal-adopting, instrumental, and egoistic 'ought'

My final point about Sidgwick's analyses is that *all* ought-judgments in the strictest sense always have two properties, that is, (1) cognition of universal truth and (2) prescriptive power that motivates a person. According to Sidgwick, this is true for any person with any kind of ethical standpoint, and it applies not only to ethical or moral judgments distinguished from prudential judgments, but also to all ought-judgments that concern an individual's practice (*ME* p. 35 ff.; Schneewind 1977, p. 226; Sidgwick 1889, p. 480).

As to ought-judgments that are usually called 'ethical', we can consider two cases. In one case, (i) a person may judge that a particular act *ought* to be done as an act of unconditional duty. In this case, the person recognizes the apparent universal truth that one ought to perform not only that particular act but also the same kind of act whenever the opportunity presents itself, and has obtained a prescription to perform that act. In the other case, (ii) a person may judge that a certain act

ought to be done in order to accomplish a certain moral end (that is, people's general happiness, or people's good in a different sense). In this case, again, he is recognizing the apparent universal truth that the proposed moral end is ultimately valid and ought to be pursued, and has obtained a prescription to pursue that end. But it is not this prescription alone which directly motivates him to perform an act which will accomplish the proposed moral end. Only when he recognizes another universal truth, that one *ought* to take the necessary or most appropriate means to that end, is he motivated to perform a particular act to attain his goal.

On the other hand, (iii) a person can also make an ought-judgment from a prudential point of view, such as when he says 'I ought to lose weight to keep my health'. This prudential judgment may be called 'ethical' in a wider sense, but more likely it will not be called so in a narrower sense. In any case, even this prudential ought-judgment contains similar features to all other practical ought-judgments, in that it demonstrates a general will that tries to control one's own future action independently of one's present desire. For the person who is swayed by his temporary impulses, such a general will is felt as a 'dictate' of reason against those impulses, and this dictate is expressed by the term 'ought'. In this case, he is, again, recognizing the apparent truth that one *ought* to pursue an end other than one's present desire (i.e., one's lifelong happiness), and has obtained the prescription to pursue that end.

However, pure egoists might claim that their lifelong happiness is not what they feel they are *prescribed* to pursue but what they *want* to pursue. If so, (iv) he will still use the term 'ought' in the way in which he says that he ought to do A in order to attain end B. This is the use of 'ought' which orders the appropriate means to an end. These 'means–end' ought-judgments do not merely show the fact that a certain means is an indispensable condition for the attainment of a given end. In making this means–end ought-judgment, a person is recognizing the universal truth that one *ought* to take the necessary or appropriate means to an end, and based on that recognition he has obtained a prescription to perform a certain act.

In all these cases from (i) to (iv), the motives urged by those prescriptions will form certain kinds of general will, and this will often conflicts with other motives or impulses. This friction is called the conflict between reason and desire.

We can see the sharpness of Sidgwick's analyses in his claim that ought-judgments are also used in egoism or in judgments of means–end conformity. Also remarkable are his insights that ought-judgments in a

teleological context will go through two steps to motivate us to action. We will later recall the following points: first, when we adopt teleology in discussing what one ought to do, we must admit not only that one ought to pursue a certain end, but also that one ought to choose an action which is a necessary or appropriate means to that end; second, even an egoist who never takes a moral point of view can make an ought-judgment. In doing so, he is either recognizing that he ought to pursue an egoistic end, or that he ought to take the proper means to achieve his egoistic end.

4.3 'Good'

Sidgwick's explanation of the concept of good appears in Book I, Chapter 9 and Book III, Chapter 14 of *ME*. In Book I Sidgwick develops a conceptual analysis of good, and in Book III he refers to this analysis in the context of his proof of hedonism. Before we go into his analysis of good, we should note a few things. First, Sidgwick attempts to define the meaning of good in such a way that it applies to all the uses of the term good. What he would like to elucidate is, however, not what is good as a means to something else, but what is good in itself, which is called an *ultimate good*. Second, Sidgwick mainly concentrates on the definition of *good on the whole*. He defines this concept with three steps: he first considers 'good for me at present', then defines 'good on the whole for me', in which one takes into account one's future self, and finally reaches the notion of 'good on the whole' which does not refer to any particular subject.

His main points concerning the concept of good are as follows. One, good is related to desire – though their relationship is more complicated than it first appears. Two, good is a comparative concept. If a good is to be compared to other goods, there has to be some common criterion to compare them. Such a criterion has to be a universal one that can be recognized by all rational beings. Third, if a certain good is good in itself and is considered as the *best* on the whole, and if it turns out that one can actualize it, that good becomes the end that we *ought* to aim at. In short, good is the concept which links desire with 'ought' via the process of comparison.

4.3.1 Findings from common usage

As usual, Sidgwick deals with the concept of 'good' in the sense we commonly use. We can see the following features regarding the common-sense concept of good.

4.3.1.1 Good is attributed to various objects including unattainable ones, and their goodness is mutually comparable

The concept of good can be applied to actions, physical objects, feelings, a person's character, and various other things. Right action is considered as a type of good action, and virtue is regarded as an aspect of good character. The concept of good can even be applied to things which are impossible for us to attain or accomplish. In addition, good is a comparative concept that allows comparisons among various objects. This means that we can compare different kinds of good and discuss which are better or not. We can also express this comparison in quantitative terms, as when we talk about a *greater* or *lesser* good. If we need to compare different goods, we need to make clear which criteria we are using to evaluate and compare their goodness.

4.3.1.2 Good as a means and the ultimate good

The term 'good' is sometimes used to describe an appropriate means to produce a certain effect. However, we sometimes judge an action itself, or the end of an action, as 'good'. Good in this context is related to the ultimate reasons for action; it is this sense of good that is the primary concern of this book. What is good in itself is called an ultimate good in contrast with what is good as a means to attain something else. Sidgwick focuses on an analysis of the concept of ultimate good (so hereafter I mean by 'good' the ultimate good, unless otherwise suggested). In doing so, Sidgwick, as usual, attempts to clarify the meaning of this concept by exploring commonsense. That is, step by step, he clarifies the meaning of the ultimate good by examining what we usually judge to be good in itself. We should note, however, that what he describes here is the *definition* of ultimate good. The extensive meaning of this concept is not yet explored.

4.3.1.3 Ultimate good relates on the one hand to desire or choice, and on the other, to what one ought to aim at

Sidgwick attempts to interpret the ultimate good in relation to desire or choice in human beings (*ME* p. 106). What is good in itself seems to relate to our desire. We commonly regard the object or satisfaction of a present desire as '*pro tanto* "a good"' (*ME* p. 381) as long as only its intrinsic value is considered and excludes its bad consequences or collateral effects. Also, when we judge something as good, we tend to desire to obtain it if possible. When we talk about people's character or acts that are good in themselves and not merely as a means to something else, we are expressing a kind of attractive moral ideal, and something which one would recommend and praise (*ME* pp. 105–6).

Nevertheless, a good does not simply mean that which we actually desire. Sidgwick suggests the possibility in which we may define 'good' or 'good acts' as 'what one ought to aim at' or 'what ought to be pursued or chosen' (*ME* pp. 377, 381). While 'good' somehow relates to our desire, it also seems to relate to the dictate of reason in that we *ought* to pursue it.

Sidgwick argues, however, that, in our judgment about good, the 'dictate of reason' is only latent and we do not feel it as an expressly authoritative command. There are several reasons for this. First, good only arouses a mild desire in the person who judges it to be good, and that person may not feel a strong binding force from making such a judgment. Second, since what we judge to be good includes unattainable things, we cannot meaningfully insist that we ought to pursue them. Third and most important, there could be many things which may be called good at the same time. Therefore, if one judges a certain thing to be good, he cannot determine whether he ought to purse that goal, because there could still be others that are judged to be good and ought to be pursued instead. It might be said, however, that, when confronted with such multiple goods, we distinguish which ones are attainable and which ones are not, and make a choice among attainable ones to determine our goal. The good we finally choose can be called the end we ought to prefer to other goods. 'Prefer' here suggests the existence of our desire for the good in question, but the phrase '*ought* to prefer' implies the dictate of reason.

The common-sense understanding of the concept of good explained above is still vague. Sidgwick further conducts philosophical analysis to clarify this concept.

4.3.2 What is judged 'the greatest good on the whole' will be the targeted end

At this point I would like to highlight a premise which Sidgwick must have assumed. This is the premise that the proposition 'what one now judges to be the *greatest* good on the whole becomes, if possible to attain, an end which one *ought* to pursue' is analytically true, owing to the meaning of the phrase 'the *greatest* good'.

In the previous section we pointed out that good is a comparative concept. We do talk about better or worse, and the greater or the lesser good. But why do we compare different types of good? It must be in order to determine the end we ought to seek.

Let us examine this point in more detail. We are here concerned with making a decision about an action that we ought to take in a particular

situation. Such an act may not necessarily be one in which a person pursues a certain end. Even if it is a kind of act that targets a certain end, that act might simply be the one that a person ought to take merely as an appropriate means to the end that he *wants* to pursue. Nevertheless, if that is the kind of action which we ought to take because it aims at a certain *ultimate good* that we *ought to* pursue, we have to determine which good is the one we ought to seek. Thus, we feel the need to compare different ultimate goods. In this context, we compare different goods in order to determine the end that we should now pursue.

In the situation stated above, if a person now judges that he *ought to* prefer one good over another, he can call the former 'what is *better* for himself at present', or to put it in quantitative terms, 'the *greater* good for him at present'. Furthermore, when he determines that he *ought to* prefer a certain good to all other goods at the time of choosing it, he can call it 'the *best*' or 'the *greatest* good'. Thus, at present the person should prefer what he now judges to be the greater good to what he judges to be the lesser good. Additionally, from a teleological point of view, a person *ought to* aim at what he presently judges to be the *greatest* good on the whole – that is, after comparing all the conceivable goods. My point here is that this is *an analytical truth* that can be derived from the very definition of the concept of better/best, lesser/greater, and the greatest good.[3]

Although Sidgwick himself does not develop a clear and definite argument about it, it is important for us to note the point that *by definition* the greatest good on the whole becomes the end that one ought to pursue, from the teleological viewpoint. This is important because herein lies a key for understanding the maximization principle of utilitarianism, which tells us to pursue people's greatest good on the whole.[4] (However, this is not the sole point of the maximization principle. We need to consider other elements in order to prove the utilitarian principle of maximizing the sum total of people's good. I will fully discuss this later.) In any case, it is evident from several passages that Sidgwick's entire argument is based on this premise. For example, Sidgwick suggests that we do think that elements of ultimate good are quantitatively comparable, and that in comparing goods we profess that we prefer a greater good to a smaller one (*ME* p. 110). It must be added, however, that the term 'prefer' is insufficient in this context. Sidgwick also claims that we need the criterion of comparison in order to decide which good we *ought* to prefer or to pursue (*ME* p. 106; *ME* Bk. 3 Ch. 14 p. 406). I think this latter point is the more precise explanation about the relationship between goodness and action. The expression that a certain

good is better, or greater in amount, than another good must be used to show that one *ought* to prefer the former good.

Of course, we cannot decide which good we ought to pursue just by comparing two different goods. If there is still another good which is greater than both, we ought to prefer it. A good that is greater than all other goods is the good which we ought to prefer to all others, and it is called 'the greatest good on the whole'. Thus, by definition, the greatest good on the whole is the one we ought to pursue, if it is attainable. I believe this is the most plausible interpretation of Sidgwick's argument about the comparison of goods and its relation to action. Indeed, Sidgwick himself states that we cannot say one ought to perform an act that brings about a certain good unless this good is judged to be the greatest good (*ME* p. 113, for example). I assume this point throughout the rest of the book.

We should note, however, the following. First, because the statements that 'one ought to prefer what one now judges to be the greater good to what one judges to be the lesser good' or that 'one ought to pursue what one now judges to be the best' are only analytically true by the definition of 'the greater good' or 'the best'; they do not provide us a substantial guidance for action. That is, even if we understand that the greatest good is by definition what we ought to pursue when possible, it does not follow from this formal principle that we know what action we actually ought to take. In order to decide what we actually ought to do, we need to determine the substantial content of the good we are seeking. Therefore, Sidgwick never claims these propositions to be fundamental moral principles. Truly significant moral principles must have substantial content, unlike the propositions that are true by definition in the way described above. Nevertheless, it is legitimate for Sidgwick to assume this proposition, unless he misunderstands it as a proposition that offers substantial guidance for action.

Second, what is posited above is an explanation about the size or amount of good for a person at present, that is, what this very person now judges to be good. It is not about what this person judges to be good *at some future point* nor about what *others* judge to be good for him. The question of how one should deal with the latter two kinds of good will be taken up later.

Third, though we have just argued that 'the greater good' amounts to 'what the person ought to prefer at the time of judging it', we have not explained anything about the question of *what criteria* that person should use to determine what he *ought* to prefer ('a person ought to prefer X' is different from 'he actually prefers X'). In other words, the

substantial criterion for the comparison of good is not yet determined. We can suggest, however, that the criterion for the comparison of good must be a universal one. This is because the term 'ought' implies a universal judgment, which all rational beings would admit regardless of differences among particulars. This is an important point.

Finally, we cannot say that the action which brings about the greatest good is the only one that we commonly judge to be the right action, or the action which ought to be done. The actions we regard as right include not only those which pursue the best end but also those which are judged to be duties in themselves. In addition, one may judge a certain action to be the best action in itself and not as a means to a good end; this is also different from what is meant by the term 'the action which ought to be done'. This is because, even if it is judged to be the best action, we cannot say that one ought (in the strictest ethical sense of the term) to do it when that action is unfeasible. Here, one may say that 'the action which is best in itself and feasible' equals to 'the action which one ought to do'. Still, right actions and good or best actions give us different impressions, in that 'right' actions often give us binding commands, whereas 'good' evokes a mild inclination in persons judging those actions to be good.

4.3.3 Good as not equivalent to pleasure

Thus, Sidgwick claims that 'good' relates, on the one hand, to ends which ought to be pursued and to desire on the other. At this point, some may be inclined to associate this claim with hedonism, which regards pleasure, that is, desirable feeling as the end one ought to seek. One may easily suspect that good has a close relationship with pleasure. In this context, Sidgwick examines the possibility of whether good can be directly *defined* as pleasure.

According to the judgments we naturally make, however, 'good' and 'pleasure' do not have the same meaning. Indeed, when we use the term 'good' for objects other than actions or people's character and when we judge them to be good in themselves – for example, when a certain dish, wine, poem, or music is recognized as good in itself – there seems to be a close correspondence between the affirmation of goodness and the pleasure derived from those objects (*ME* p. 106), but on close examination we can find significant differences between pleasure and goodness.

First, though we often judge an object to be good because it gives us pleasure, such pleasure is usually limited to a specific kind, and even if the same object gives us pleasure of a different kind, we seldom think

such different types of pleasure constitute a reason for calling that object good. Even if, for example, a painting gives its seller an enormous amount of pleasure because it fetches an extremely high price, we do not think that this picture is good *solely because* it gives him such monetary delight. We usually consider a painting good because it gives us aesthetic pleasure. The same is true of judgments about good acts. According to Sidgwick, our common judgment on good action is somewhat analogous to the perception of beauty in material objects. We make such a judgment when the conduct causes 'contemplative satisfaction [. . .] to a disinterested spectator' (*ME* p. 109). Even if other kinds of concomitant pleasure are caused by the same conduct, they are irrelevant to the judgment of the goodness of that conduct. We usually call an act of veracity good because it gives pleasant impressions to the persons who witness that act. Whether this act later influences someone's pleasure or pain is often irrelevant to such an evaluation. Thus, as far as our ordinary judgments are concerned, goodness does not correspond to all kinds of pleasure. Second, we usually think that, as regards an individual's pleasure, the person who is feeling it is the final judge, but, when we judge the goodness of a certain object, we presuppose some universally valid criterion, apart from a mere personal view or feeling. The person who can properly use this criterion is called a man of good taste, and many believe that only the judgments of such a man constitute the true evaluation of the goodness of things. Additionally, connoisseurs of wine or of works of art may be able to properly judge the goodness of objects even when they do not enjoy them. It is also said that those who have particularly rich feelings may derive greater pleasure than the pleasure others obtain from 'better' objects. Thus, the pleasure obtained from a certain object is not necessarily in proportion to the goodness of the object. Third and finally, if the claim that pleasure is the ultimate good is, as many philosophers have asserted, to be admitted as a nontautological and truly significant one, good is not just a synonym for pleasure.

In sum, although 'good' things certainly frequently correspond to what are 'pleasurable', their meanings are not the same; this only means that the actual objects these terms indicate are often the same (*ME* p. 109).

4.3.4 Not what is desired, but what is 'desirable'

Then, how about interpreting 'good' more plainly, as being equivalent to the object of desire? As we have already discussed, however, good is not what we simply desire. Good is certainly related to desire in some

way, but it also involves the latent dictate of reason that 'one ought to pursue it if it is the greatest attainable good'.

Sidgwick's own view is as follows. According to him, good is not necessarily the object which one actually desires, but the object which one *would* desire at present if ideal conditions were satisfied. Thus, he explains good in terms of what is 'desirable', which is distinguished from what is actually desired.

> It would seem then, that if we interpret the notion 'good' in relation to 'desire', we must identify it not with the actually *desired*, but rather the *desirable*: – meaning by 'desirable' not necessarily 'what *ought* to be desired' but what would be desired, with strength proportioned to the degree of desirability, if it were judged attainable by voluntary action, supposing the desirer to possess a perfect forecast, emotional as well as intellectual, of the state of attainment or fruition.
>
> (*ME* pp. 110–1. Italics as in the original text.)

Here, Sidgwick points out that what is 'desirable' does not necessarily mean what ought to be desired. This is because the use of 'ought' would suggest an explicit dictate of reason, which is incongruous with the concept of good, the dictate of which is allegedly latent. As explained before, the concept of good does not have an explicit binding power, but arouse a mild desire which contains certain ideal elements. Also, the concept of good cannot impose an explicit dictate that one must aim at the good in question when one can only imagine the situation in which it can be attained but one cannot actually attain it. Additionally, the existence of a good does not entail an explicit order to pursue it if there is another good which ought to be preferred. Still, what a person judges to be the greatest good on the whole, after comparing and balancing all attainable goods, becomes the end that he *ought* to aim at. In addition, when a certain good that is attainable for him is adopted as an end of his action, the act that is a necessary or appropriate means to attain that end becomes what he *ought* to do (*ME* p. 112 fn. 1; *ME* Bk. 1 Ch. 3 Sec. 4). In this sense, good is implicitly related to what one ought to do.

Then, what is the criterion for comparing different types of good? Here, we find, in the passage cited above, the phrase 'what would be desired, with strength proportioned to the degree of desirability'. According to Sidgwick, good is what would be desired under certain ideal conditions, and the degree of its desirability varies among different goods. Some good would be strongly desired on those conditions, while others would be less strongly desired. At this point, we may be

inclined to conclude that this degree of desirability, that is, the strength of desire that occurs under ideal conditions, can serve as the criterion for comparing different goods. Under the ideal conditions described before, a person would desire a good that is most strongly desired among all goods (this proposition is analytically true), and would be motivated to adopt it as his end. *If* we could regard such a good as 'the greatest good', the good that one would most strongly desire under the alleged ideal conditions would be the good to be aimed at. However, Sidgwick never adopts this line of argument. The criterion for comparing goods is a criterion to decide which good ought to be preferred, and such a criterion must be a universal criterion that all rational beings would accept. Thus, for Sidgwick, that *an individual* would strongly desire a certain good under ideal conditions is not enough to serve as such a *universal* criterion. In Book I, Chapter 9 of *ME*, where he analyzes the concept of good, Sidgwick states that, 'It remains to consider by what standard the value of conduct or character, thus intuitively judged to be good in itself, is to be co-ordinated and compared with that of other good things. I shall not now attempt to establish such a standard' (*ME* p. 113). He does not attempt to establish such a criterion at this point because the substantial content of good has not yet been determined. Later in *ME* (Bk. 3 Ch. 14) Sidgwick proves the universal truth that the substantial content of ultimate good is pleasure, and then he finally determines the criterion for comparing different good things.

Now we have seen that a good is what is desirable, and that the greatest attainable good on the whole becomes, by definition, the end that ought to be pursued. This analysis is not yet complete, however. As explained above, what one ought to do is concerned not with a mere good, but the *greatest* good *on the whole*, which becomes the end one ought to pursue. We have explained the meaning of the adjective 'the greatest', but have not yet fully interpreted the phrase 'on the whole'. I suggested that 'the good on the whole' is what is judged to be good after comparing and balancing all good things. I have not explained, however, how to balance them.

Thus, Sidgwick next considers how to judge 'the good on the whole'. In doing so, he defines three stages of good, which are 'good for me at present', 'good on the whole for me', and 'the good on the whole'.

4.3.5 Good for me at present

We commonly regard the object or the satisfaction of one's own present desire as 'a good' as far as it is considered by itself and without taking into account its collateral or subsequent effects or any other external

factors. Sidgwick starts by considering 'an ultimate good for me at present' before examining 'the good on the whole' because it is with the former that we can most clearly observe the relationship among desire, good, and choice. Sidgwick himself calls it 'his [one's] own Good and ultimate Good' (*ME* p. 109), but I believe it is more precise to call it 'an ultimate good for me at present'. This means what a person judges to be good for himself when he only considers the state of its fulfillment and never considers any other consequences. This good appears to be closely connected to his present desire. Therefore, Sidgwick tentatively explains it as 'what a man desires for itself – not as a means to an ulterior result – and for himself – not benevolently for others' (*ME* p. 109).

However, as I previously discussed, even when we only consider the state of its fulfillment, what we now desire is not always our true good. What one desires at present is but an apparent good. This is so because, when the object of desire is obtained (in other words, when the desire is satisfied), it may turn out that it is not good, or not as good as expected. In addition, a prudent person, while he is strongly inclined to desire a certain object, may suppress his desire because he knows he cannot obtain it by his own effort (*ME* p. 110). A good for me at present is not what I desire at present, but what is desirable for me at present. This is what I *would* desire *in itself* – and not as a means to any ulterior result – and *for myself* – and not for others *if* (1) I judge it possible to attain and *if* (2) I fully expect the state of attainment when my desire for it is fulfilled.

4.3.6 Ultimate good on the whole for me

Then, is 'the good *for me at present*', in the sense described above, equivalent to what *I presently judge* to be good on the whole? Sidgwick opposes this idea. The good for me at present, that is, what I would now desire in itself and for myself when it is attainable and when I fully imagine its fulfillment, may not be good for me on the whole if I consider its collateral and subsequent effects. Good for me at present may have unfavorable consequences for me in the future, when I might have a different desire from the present one. In addition, what I do now may change my way of life in the future. I may prevent my future self from pursuing certain things, including what I might have judged to be desirable. Suppose, for example, I indulge in drinking. I start drinking because at a certain point of time (t_1), I judge it to be desirable in itself and for myself, just by imagining the satisfaction of my desire for alcohol, while I never think of its consequences at a later point of time (t_n). Then I keep drinking and finally become an alcoholic. The drunken me

at t_n no longer desires an academic life or even a normal healthy life. In this case, I at t_1 have made my future self at t_n into an alcoholic and hence no longer desire to perform academic research, which I would have judged to be the most desirable thing in life. We would therefore conclude that my choice at t_1 is not good on the whole for me.

With this consideration, Sidgwick provides the following definition of 'my future good on the whole'. He identifies this type of good as one's good on the whole, which means what is good for him from a time-neutral perspective.

> [M]an's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realised in imagination at the present point of time.
>
> (*ME* pp. 111–12)

This type of good is also what a person *now* judges to be good for himself. In order to determine this time-indifferent good, however, a person has to expect all the alternatives open to him and their consequences, to consider the possibility that his desire may change in the future, and to imagine what his future self would judge to be good for him at that point. One thing should be noted here. Sidgwick hereby provides the definition of a person's good on the whole, but he has not yet stated how to treat the good for one's future self and how one should assimilate it into one's present judgment. Here, we can only explain that what I now judge to be 'good on the whole for me' is what I would now desire if I *somehow* consider the object or the satisfaction of both my *present* and *future* preferences. It will become clear later that this definition of one's good on the whole is closely related to Sidgwick's Principle of Prudence, which states that one ought to treat one's future good equally to one's present good. At this stage, however, such a moral principle, which tells us how we *ought* to treat different goods, is not yet formally stipulated – though in Book I, Chapter 9 of *ME* Sidgwick already hints at the point that one ought to equally consider consciousness at all points in time. As far as the definition of one's good on the whole stated above is concerned, there still remains the possibility that a person may consider the good for his future self but may not give it the same weight as the good for him at present.

According to the definition stated above, the concept of good can be fully explained by describing actual and hypothetical states of affairs – by a combination of the expectation based on facts and the

desire aroused by such an expectation. Sidgwick, however, is not content with this descriptive definition of good. This is because it is commonly thought that the concept of 'good on the whole' latently implies rational prescription, which cannot be reduced into the mere description of facts.

> It seems to me, however, more in accordance with common sense to recognise – as Butler does – that the calm desire for my 'good on the whole' is *authoritative*; and therefore carries with it implicitly a rational dictate to aim at this end, if in any case a conflicting desire urges the will in an opposite direction.
>
> (*ME* p. 112. Italics as in the original text.)

Thus Sidgwick presents another definition of good, as follows:

> [W]e may keep the notion of 'dictate' or 'imperative' merely implicit and latent, – as it seems to be in ordinary judgments as to 'my good' and its opposite – by interpreting 'ultimate good on the whole for me' to mean what I should practically desire if my desires were in harmony with reason, assuming my own existence alone to be considered.[5]
>
> (*ME* p. 112)

By introducing the phrase 'if my desire were in harmony with reason', Sidgwick expresses the rational implication of the concept of 'ultimate good on the whole for me'. This means that this type of good gives a person, if he has obtained a certain cognition, a kind of motivational power which is distinct from a mere desire that this person actually has. It is his reason that recognizes that a certain alternative is good, after precisely anticipating all the consequences of all alternatives open to him. This person believes his choice is based on 'an apparent universal truth', which is independent of those particular alternatives or consequences.

We should note, however, that Sidgwick, in the second definition just quoted, never states that the good on the whole for me is *the only thing* that a person would desire to attain as a rational being. A good for me on the whole does not immediately mean the *greatest* good for me. There might be any number of good things, which a rational being thinks desirable on the whole for oneself. A certain 'good on the whole for a person' may be preferable to him than another 'good on the whole for him', in which case it is said that this person should seek the former

good. This is the reason why the definition shown above only latently contains the dictate of reason. Nevertheless, if I find the *greatest* good on the whole for me, I usually regard it as what I ought to pursue for myself. It is at this point when the dictate of reason becomes explicit.

Thus, Sidgwick explains good as what is 'reasonably desired', 'that at which it is ultimately reasonable to aim' and 'the ultimate end of rational conduct' (Contents of *ME*7, Bk. 1 Ch. 9 Sec. 3; *ME*7 pp. 3, 91 and p. 92 fn. 1; *ME*1 p. 93). This means that the concept of good contains the potential dictate of reason, in the sense that it has a motivational power derived from reason, and that it becomes the end which ought to be pursued when it turns out to be the greatest good.

4.3.7 Ultimate good on the whole

Simply defining 'the ultimate good on the whole for me' is not the end of the story, however. What I at present judge to be good is not necessarily what is good on the whole *for me*. We often judge a certain act or character as good in itself, whether it relates to some particular individual or not. For example, when we talk about a veracious person, we often praise him not because he benefits us, but because such a character appears to be good in itself. In making such a judgment, we do not mean that it is good for a *particular* individual (whether this individual is myself or someone else), but that it is considered to be good regardless of its contribution to some particular individual. In this sense, we are judging what is good from an impartial perspective – in other words, we are hereby forming the notion of 'the ultimate good on the whole' without referring to a particular subject. Sidgwick defines this as follows:

> '[U]ltimate good on the whole', unqualified by reference to a particular subject, must be taken to mean what as a rational being I should desire and seek to realise, assuming myself to have an equal concern for all existence.
>
> (*ME* p. 112)

As in the previous explanation of good, the good defined here means what *I at present* would desire under ideal conditions, that is, *if* I *were* a rational subject, representing every detail of the situation at present and in the future, and gave equal consideration to everything involved. Thus, we can apply to this notion the term 'desirable' in the sense previously stated. Besides, by using the phrase 'a rational being', this definition also expresses recognition of truth and motivation. What reason apprehends here is the universal truth, '*this* is the universal good', which one ascertains

after paying attention to all existence. This good on the whole is, however, not necessarily the *greatest* good on the whole. It should be noted that the definition shown does not imply that it is the *only* thing that as a rational being I should desire and seek to realize. Even when I judge something to be good on the whole, there might be a *better* good on the whole, which I might then consider pursuing. Sidgwick himself argues as follows, immediately after presenting the definition just stated.

> Such a judgment differs, as I have said, from the judgment that conduct is 'right,' in so far as it does not involve a definite precept to perform it; since it still leaves it an open question whether *this particular kind* of good is the greatest good that we can under the circumstances obtain.
>
> (*ME* p. 113. Emphasis added.)

For this reason, we cannot equate the ultimate good on the whole with what we ought to aim for. When this good on the whole proves to be the *greatest attainable* good, however, it becomes the end that one *ought* to seek. In this sense, 'the ultimate good on the whole' latently contains the dictate of reason.

Now we have finally reached the definition of 'the good on the whole' a subject judges at present. We should remember, however, that a phrase in this final definition, 'to have an equal concern for all existence', does *not* yet imply that I should *put equal weight* to the good of every being. Another thing to be emphasized is that at this point there is no indication that this 'good on the whole' can or should be explained in terms of the *sum total* of the good of individuals. To put it another way, there is still a gap between having an equal concern for all and counting the good of all beings *according to their size*. A similar gap exists between what a rational being would aim for and what is the greatest *sum total* of such good. When Sidgwick explains the method of utilitarianism in Book IV of *ME*, however, he appropriates the egoistic method of evaluating individual pleasure and pain for assessing universal happiness. In doing this, Sidgwick obviously assumes that people's general happiness (which corresponds to 'the good on the whole' in utilitarianism) is the aggregation of pleasure for each individual (this latter pleasure corresponds to the good for each individual).[6] The idea that the good on the whole is the *sum total* of the good for each individual will be supported later by three steps: the two definitions of 'good on the whole' just described, two intuitive principles that state how good should be treated, and the proof of hedonism. This will become clear later in 7.1 and 7.3 of the present book.

4.4 Relationship between 'right/ought' and 'good'

I believe that the previous exposition has already suggested how the notions of 'right/ought' and 'good' are related to each other. The terms 'right' and 'ought' are used to guide an action, and for that reason they are the notions central to ethical consideration. The notion of 'good' is related to our desire, and is applied not only to actions but also to various other entities, including those that are impossible for us to attain. It should also be noted that 'good' is essentially a comparative notion. Thus, the judgment that a certain act is 'right' contains an authoritative prescription to perform that act, whereas the judgment that a certain act or object is 'good' does not accompany the dictate to perform that act or to attain that object *unless* it becomes clear that such an act or object is the greatest attainable good in those circumstances and that such an act or object can be realized by our voluntary effort. On this account 'good' and 'right' are different.

However, these notions are similar in that both have rational implications. This means that each of these notions explicitly or implicitly contains (1) the recognition of a certain universal truth, and (2) a dictate based on such recognition. As far as we reason within a teleological frame, the greatest attainable good becomes an end that one ought to seek, and, when we adopt a certain good as the end of action, the action which is necessary to attain that good, or is the most appropriate means to that end, becomes what *ought* to be performed. Thus, under certain conditions, the notion of good can also imply the dictate of reason, expressed as 'one *ought* to seek this' or 'one *ought* to perform a particular type of action to attain this end'.

The first edition of *ME* clearly stated that the right action must be the best attainable action (*ME*1 p. 99; *ME*2 p. 100; Schneewind 1977, pp. 225–6). This statement, however, is insufficient. To be precise, not all right actions are the actions that pursue the best end. There are other types of right actions, that is, the actions that are simply done out of a sense of duty, in which cases any 'good' (the end of action) is perceived by the agent at the time of execution. Furthermore, although it is true that the best attainable action involves the dictate of reason, such a mandate does not appear compulsory or binding, but only mildly elicits desire in a person who acknowledges the goodness of the action.

Still, it will later become clear that we need to determine good in order to determine the ultimately true and right action. When we closely analyze common-sense duties and virtues, we come to realize that to determine right actions we need some fundamental principles *other than* the common rules of duty and virtue, and, when we clarify

those fundamental principles, we notice that they inevitably involve the notion of good. I will elucidate this point in the next chapter.

4.5 Good and human consciousness

One more thing should be noted here. Sidgwick presents another argument at the end of his analysis of the notion of good. He claims that, although we usually judge various things to be good, if we reflect on what is permanently judged to be good, nothing can be regarded as good – good in the sense of what ought to be sought – apart from its relation to human existence, or to some consciousness or feeling (*ME* Bk. 1 Ch. 9 Sec. 4 p. 113). For instance, we often judge some beautiful lifeless object to be good, but we do not think it rational to produce such beauty in remote places where humans can never contemplate it. Sidgwick goes further to claim that beauty, knowledge, and other abstract good things can be rationally sought by humans *insofar as* they contribute either to happiness, or to the perfection or the excellence of human existence. (As I previously explained, 'rationally sought' is the expression that is used when it is still uncertain whether the object really ought to be sought, but that implies the *potential* for something to become what ought to be sought.) Certainly, when a person is creating some beautiful object, trying to realize his ideal, or pursuing knowledge, he is often absorbed in the task before him, without thinking about its influence on himself or on other humans. Sidgwick claims, however, that when we reflect on why such creation or pursuit is significant for him, and when we clearly understand all the possible answers to that question, we would admit the significance of his action only when we understand that such creation or pursuit will bring happiness, excellence, or perfection to someone. There are two exceptions to this statement, however. First, as for happiness, we can include the pleasure of nonhuman animals into the proper ends of our conduct, but Sidgwick says we usually do not aim for 'perfection' in nonhuman animals, except as a means to our ends or as objects of our scientific or aesthetic contemplation. Therefore, perfecting someone is always construed as perfecting human existence. Second, Sidgwick also claims that happiness or perfection of such superhuman beings as God cannot become our practical end, and therefore that such beings should be excluded from our consideration. Thus, Sidgwick concludes at the end of Book I of *ME* that 'if there be any Good other than Happiness to be sought by man, as an ultimate practical end, it can only be the Goodness, Perfection, or Excellence of Human Existence' (*ME* p. 115). This is one of the crucial arguments that will later be used in his proof of hedonism.

5
Testing the Significance of Apparent Truths

After finishing a series of examination of common-sense morality, in Book III, Chapter 11, Section 2 of *ME*, Sidgwick presents 'four conditions, the complete fulfilment of which would establish a significant proposition, apparently self-evident, in the highest degree of certainty attainable: and which must be approximately realised by the premises of our reasoning in any inquiry, if that reasoning is to lead us cogently to trustworthy conclusions' (*ME* p. 338).[1]

The phrase 'an apparently self-evident proposition' means that it is intuitively known to be true. When something is self-evident, its truth is known without any need of further reasoning or demonstration.[2] Sidgwick regards, however, what is *tautologically* self-evident as useless for our knowledge and excludes it from his argument here. This is because Sidgwick presents the conditions that propositions must satisfy to give us a practical guide when we deal with ethical questions. Any practical inference must assume these conditions if it intends to lead to a persuasive conclusion. A 'significant' proposition implies that it is not merely tautological, but it provides us with a certain substantial guide.

Sidgwick claims that these four conditions are used to establish such a proposition 'in the highest degree of certainty attainable' – if not with absolute certainty. This suggests that these four conditions may not be exhaustive, but they are all the conditions that we can imagine, for establishing self-evident and significant propositions.

To restate this: Sidgwick presents four necessary conditions for establishing a 'significant' intuitive proposition – 'significant' in the sense that it gives us a useful guide to solve our questions. These conditions must be assumed by any practical inference that aims to solve our questions.

Sidgwick's own intention here is to ascertain that common rules of duties and virtues cannot be regarded as self-evident and significant

axioms in the light of these conditions. By doing this, Sidgwick shows the limit of dogmatic intuitionism and recommends us to proceed to philosophical intuitionism. Nevertheless, these four conditions must also be met when we explore true philosophical intuitions which lead to three fundamental principles, as long as we use rational reasoning in the process. I will articulate these four conditions that make up the most important presupposition of Sidgwick's arguments, and one additional condition that I believe is also important to understand the structure of his argument.

5.1 Four conditions proposed by Sidgwick

The four conditions of 'self-evident and significant propositions' are stated as follows:

1. That the terms of the proposition be clear and precise.
2. That the self-evidence of the proposition be ascertained by careful reflection.
3. That the propositions accepted as self-evident be mutually consistent.
4. That the propositions be supported by 'universal' or 'general' consent among people – especially among the experts who are familiar with the matter in question.

It is said that these conditions are derived from the work of a Scottish intuitionist Thomas Reid (1710–96), and it is possible to assume that Sidgwick simply followed the philosophical trend of his time.[3] Sidgwick still explains why each of these conditions should be adopted. We need to examine his arguments in the following sections.

5.1.1 Clarity and precision of the terms

Sidgwick makes only a few remarks on the first condition – that the terms of propositions be clear and precise. He simply points out that the two rival originators of modern methodology, Descartes and Bacon, commonly stressed the importance of this condition, and that Bacon's warning against wrongly fixed notions in our common thinking (*notiones male terminatae*) is especially to be noted in ethical discussions. We can reconstruct, however, the point that Sidgwick wants to make, from this brief passage and from his previous arguments in *ME*.

First, when we are to deal with an ethical question, we naturally become involved in a thinking process – and we attempt to think logically if we seriously want to solve the question. As long as our

thought is logical, its content takes the form of a proposition using several terms. For example, when we deal with the ethical issue of abortion, we are considering if 'one ought not to kill a human embryo' or if 'it is a bad thing to abort an unborn child', etc. We are asking whether that proposition is self-evident or not, or if it contains truth or not. This being the case, we must clearly understand the proposition that we are considering, and the terms used in it (for example, 'ought', 'kill', 'human embryo', and so on), before we can successfully answer our question. That is, the meaning of each term must either be clear to us or be defined in different terms that are clear to us. For example, if we do not know what 'a human embryo' is, we may understand it by analyzing this term as 'a being at the early stage of human development', etc. Especially in ethical discussions, where we are trying to reach consensus on establishing common rules of conduct, all of us must clearly understand the proposition that we are discussing, and the terms that compose it. Thus, Sidgwick believes that no one would object to this condition of terminological clarity being uniformly imposed on us. He is further convinced of its validity by the fact that even the two great thinkers, who often severely disagreed with each other, agreed on this condition as the basis of their philosophical method.

Needless to say, it will not suffice if the terms are *apparently* clear and precise to ordinary people. The terms used must be truly clear, even to people who seriously reflect on the matter. Then, how can we judge which terms are truly clear and which are not? Sidgwick states that 'In fact my chief business in the preceding survey has been to free the common terms of Ethics, as far as possible, from objection on this score' (*ME* p. 339). This probably refers to his efforts in the previous chapters in *ME* to clarify the meaning of terms such as 'right', 'ought', 'good', and 'intuition', and to his definitions of each particular virtue in the explanation of common-sense morality. Sidgwick certainly believes that these kinds of terms never become clear unless accurately defined. One common feature of most of these terms is that they are not terms that merely describe observable facts, but are those that contain, or relate to, evaluative judgments. An evaluative judgment usually depends on a person's taste or other subjective factors, and therefore people's understanding of such evaluative terms can differ much more greatly than that of merely descriptive terms. Thus, we can say that a term that contains or is related to certain evaluative judgments cannot be clear unless its meaning is precisely analyzed and defined.

There are, however, several notions, such as 'right' or 'ought', which are too fundamental to be formally defined. We can only clarify the

meaning of such indefinable notions or terms by precisely determining their relationship to other notions, or by pointing out the properties of the judgments that contain those terms (see *ME* pp. 32–3).

We can also get an idea of which terms Sidgwick believes are clear or unclear by looking over some passages in his examination of common-sense morality, where he labels several terms as ambiguous. For example, the term 'freedom' is ambiguous (*ME* p. 275). This term is ambiguous because 'freedom' can sometimes be used to mean a situation in which any action can be done without constraint, but at other times it can also mean a situation in which one is exempt from pain or annoyance inflicted by others. In another example, we commonly regard it a duty to appreciate another person's kindness, and we may analyze this duty as a duty of equal requital to our benefactor. But the term 'equal' here is ambiguous, because we cannot clearly decide whether it denotes compensation that is in proportion to the benefactor's *effort* or to the *benefit* that we received (*ME* p. 261). Judging from these arguments, terms that allow multiple interpretations are ambiguous and therefore unclear.

Thus, we can point out the following: particular evaluative and/or equivocal words are unclear, unless they are defined by simpler terms, or their relationship to other concepts is precisely determined, so that they are no longer understood by people differently.

5.1.2 Conviction on reflection

The second condition, 'the self-evidence of the proposition be ascertained by careful reflection', is a necessary one, because an individual's intuition does not always properly apprehend the truth, though it may *appear* to perceive a certain significant proposition as 'apparent' truth. Especially in ethics, which is related to one's motive, will, and action, our judgments are susceptible to our own strong feeling or desire, or to the opinions of others or those of the general public. We tend to judge what we *now* desire to be *desirable*, and we are tempted to approve actions that will bring about the most pleasure to ourselves. Our minds are often subconsciously influenced by such external authorities as convention, law, or tradition, so that we may easily accept widely recognized assumptions without sufficient grounds (*ME* pp. 339–41). Because of such mental weakness, we have to carefully reflect on whether what we now believe as apparent truth is really true, or merely *falsely* believed to be true because of such influences.

Sidgwick is not very clear about how to undertake this reflection. In considering the above argument, however, one such reflective method would be to ask if we have a particularly strong emotion or desire, or

whether convention, law, or tradition is affecting our present judgment, and to consider if we can still be convinced that the proposition in question is true when we put aside those affective factors.

It is certain that we cannot be completely exempt from fallacy with such reflection (*ME* p. 339). But we cannot trust a proposition whose apparent self-evidence vacillates under this kind of reflection, unless further reflection shows this proposition is indeed true. Truly self-evident propositions, however, will not be easily overturned by such examination. Thus, Sidgwick considers this second condition to be a necessary one in order to ascertain the genuine self-evidence of apparently true propositions.

5.1.3 Consistency

As for the condition that propositions accepted as self-evident be mutually consistent, Sidgwick gives the following explication:

> Here, again, it is obvious that any collision between two intuitions is a proof that there is error in one or the other, or in both [. . .] Whereas such a collision is absolute proof that at least one of the formulae needs qualification: and suggests a doubt whether the correctly qualified proposition will present itself with the same self-evidence the simpler but inadequate one; and whether we have not mistaken for an ultimate and independent axiom one that is really derivative and subordinate.
>
> (*ME* p. 341)

For Sidgwick, consistency is an absolutely necessary condition, which is based on almost the same idea as what was called the fundamental postulate of ethics (see 2.1.2 of the present book). Ethics is a systematic study that requires logical reasoning about ethical questions. In so doing, two mutually contradictory propositions cannot be true at the same time.

Such logical contradiction can also bring about practical problems. If two mutually inconsistent actions are ordered by two intuitively rational propositions, we cannot decide which action we ought to perform. If ethics is to be rational and to give us a decisive guide for what we ought to do, we have to proceed with this condition of consistency.[4]

5.1.4 Universal or general consensus

In the main text of *ME*, Sidgwick expresses the fourth condition as universal or general consent among people, but in the table of contents of

ME he rephrases this condition as the condition that propositions be 'supported by an adequate "consensus of experts"' (*ME* p. xxxiv).

The following explanation is made in the main text.

> Since it is implied in the very notion of Truth that it is essentially the same for all minds, the denial by another of a proposition that I have affirmed has a tendency to impair my confidence in its validity. And in fact 'universal' or 'general' consent has often been held to constitute by itself a sufficient evidence of the truth of the most important beliefs; and is practically the only evidence upon which the greater part of mankind can rely. A proposition accepted as true upon this ground alone has, of course, neither self-evidence nor demonstrative evidence for the mind that so accepts it; still, the secure acceptance that we commonly give to the generalisations of the empirical sciences rests [. . .] largely on the belief that other experts have seen for themselves the evidence for these generalisations, and do not materially disagree as to its adequacy. And it will be easily seen that the absence of such disagreement must remain an indispensable negative condition of the certainty of our beliefs.
>
> (*ME* pp. 341–2)

If an intuitive proposition is to be true, it must be so to everyone's mind. Therefore, ideally there should be a universal consensus about its truth. However, our judgments can be susceptible to our own feelings or to the conventions of our society, and it is even possible that all members of a society believe what is actually false to be true because all of them are similarly influenced by social convention and tradition. This is the reason why general consensus in a society cannot be sufficient proof for a proposition to be genuinely true. However, if experts who seriously examine the truthfulness of the proposition in question and who reflect on it from various perspectives unanimously support its validity, the probability of its being true seems to be sufficiently high. Or, when such experts object to a certain proposition that one has believed to be true, its truth seems to be open to doubt; one may rightly suspect that it is not the experts' judgments but one's own judgment that is mistaken. Therefore, we should keep this 'negative' condition in mind.

5.2 An additional nontautological condition

In addition to the four conditions stated above, we should remember one more condition, which is that the proposition should not be merely

tautological. Sidgwick repeatedly suggests this condition, but does not clearly state it before Book III, Chapter 13, Section 2 of *ME*. As far as we judge from Sidgwick's descriptions, what he means by 'tautological propositions' are those that can be known to be true by definition or by logic. Such tautological propositions include literally redundant ones such as 'A is A', as well as those that are analytically true by definition of the terms used. As explained earlier, we enter into the field of ethics not only to define an ethical question but to reach its solution based on sound reasoning. Therefore, we look for the kinds of propositions – precepts, advice, principles, etc. – that can serve as substantial guides for practical reasoning. Tautological propositions are certainly self-evident because they are always tautologically true, but they cannot be called 'significant' because they do not offer a practical guide to answer questions about what we ought to do.

Examples of tautological propositions are as follows:

1. To act rationally is always right (see *ME* Bk. 3 Ch. 13 Sec. 1).
2. I ought to seek what I judge now to be the greatest attainable good on the whole.

Proposition (1) gives us no practical instruction. As previously analyzed, a 'rational' action means an action that is judged to be *right* after we recognize a certain truth. Thus, proposition (1) only states that an action that is judged to be right after recognition of truth is always right. This statement makes full sense in that it correctly shows that the phrase 'to act rationally' is to be equated with 'to act right' because of the definition of the term 'rational'. However, it gives us no substantial guide to decide which truth we ought to apprehend and which action to perform, and therefore this proposition cannot be called significant.

Now, proposition (2) was previously clarified in my analysis of the concept of good. This is also a tautological proposition, being analytically true by the definition of the term 'the greatest good'. This is also an insignificant proposition in that it does not give us a substantial guide to determine the crucial question of what good we should aim for. Here, however, we should not draw the hasty conclusion that tautological propositions are always insignificant and therefore cannot serve as a basic assumption of our arguments. It is legitimate for us to proceed on the assumption that proposition (2) is true, insofar as we are aware that (2) is only true *by definition*. Sidgwick states that such propositions 'can only be defended from the charge of tautology, *if they are understood as definitions of the problem to be solved*, and not as attempts at its solution'

(*ME* p. 376. Emphasis added). What we should avoid is to be misled into thinking that these tautological propositions can give us a substantial guide for *solving* our questions.

Ethics is the study in which we attempt to obtain a substantial guide to solve the practical question of how to decide what we ought to do. If we are to obtain certain significant axioms or principles that can give us such a guide, we have to look for propositions that are not merely tautological.

5.3 The limit of common-sense morality

With the above conditions in mind, Sidgwick reexamines widely accepted maxims of common-sense morality, one by one, and shows that they cannot be truly self-evident axioms. For example, the precept of Wisdom, or the maxim that one ought to act wisely, turns out to be merely tautological and insignificant if it means that 'one always ought to do what one believes to be rational, without succumbing to an impulse that goes against it'. This is so because, according to Sidgwick, rational acts mean acts that ought to be done. For wisdom to be a significant axiom, we might interpret it as the precept that 'one ought to acquire the habit of always consciously acting rationally'. We know, however, that we sometimes attain a rational end better when we do not consciously aim for it. According to this knowledge, the precept that we ought to acquire such a habit is not self-evident at all. This contradicts the condition of 'conviction of self-evidence on reflection'. As for other common moral rules, such as promise-keeping, truth-telling, or distributive justice, they also turn out to conflict with one or more of the above conditions for self-evident and significant propositions. Their meaning cannot become clear even after we carefully analyze terms such as promise or lie. On reflection, it turns out that these precepts have exceptions or limitations, but that there is no agreement on the line between what are exceptions and what are not. It may turn out that two principles, both of which are usually accepted as self-evident, come into conflict and we cannot decide which principle is 'more' self-evident. Taking the case of debates over distributive justice, we cannot determine which principle is better in terms of distributive justice, the equal distribution principle or the principle of distribution according to merit.

In exploring ethics, in which we ask for a rational decision-making procedure of what ought to be done, we are looking for truly valid principles that can systematize our action. However, Sidgwick concludes

that we cannot find such principles among existing rules of common-sense morality. Thus, Sidgwick shifts away from scouting about for the supreme moral principle among common-sense moralities, and attempts to find truly self-evident and significant practical principles, which would satisfy the above four conditions, on philosophical reflection. Now that we learned such principles cannot be expressed in the form of existing concrete rules of conduct, Sidgwick claims that such fundamental moral principles are more abstract and can only be attained by philosophical inquiry.

> There are certain absolute practical principles, the truth of which, when they are explicitly stated, is manifest; but they are of too abstract a nature, and too universal in their scope, to enable us to ascertain by immediate application of them what we ought to do in any particular case; particular duties have still to be determined by some other method.
>
> (*ME* Bk. 3 Ch. 13 p. 379)

As such abstract principles apprehended by his philosophical intuition, Sidgwick presents his three fundamental principles of ethics, namely, the Principle of Justice, the Principle of Rational Self-Love, and the Principle of Rational Benevolence.

6
The Three Fundamental Principles

These three fundamental principles are called 'real ethical axioms – intuitive propositions of real clearness and certainty' (*ME* Bk. 3 Ch. 13 p. 373), 'self-evident moral principles of real significance' (*ME* p. 379), or 'absolute practical principles' (*ibid.*). They are principles which are apprehended by philosophical intuition, and which will provide guidance for actions that ought to be done.

These principles are different from those 'principles' which were described in 3.1.4 of the present book, where we identified three methods of ethics, namely, the propositions that identified ultimate common reasons for action, such as that 'one ought to seek one's own happiness', 'one ought to seek people's general happiness', or 'one ought to follow the rules of duty or virtue'. The latter principles are *prima facie* principles for action that can be obtained by observing what kind of policies people use to guide their action. It remains an open question as to whether these were truly valid principles.

In contrast, the three fundamental principles now under consideration are axioms that are apprehended by philosophical intuition and therefore should be admitted by all rational beings as truly valid. These principles provide the real foundation of ethics.

Three such fundamental principles, that of Justice, Rational Self-Love, and Rational Benevolence, are often called 'maxims', for they are not principles that simply state facts, but are those that serve as a guide for determining actions that ought to be done.

These principles are apprehended by philosophical intuition. As explained earlier, 'apprehended by intuition' means that some apparent truth is immediately known, apart from any induction from experiences or other inferences (see 3.2 of the present book). Thus, these principles are not supposed to be logically derived from other premises. Furthermore,

they are abstract principles, which cannot be obtained just by reformulating common moral rules in a more refined manner. For such reasons, the three fundamental principles are somewhat abruptly presented in Book III, Chapter 13, Section 3 of *ME*, with few explanations about why and how these principles came about or were stumbled on. Sidgwick focuses on precisely describing each principle rather than examining it. (As I already emphasized, to intuitively apprehend a principle means to know it as self-evident without demonstration; it does not follow that its precise formulation immediately occurs to us. However, we can understand its validity without any reasoning once we are presented the principle precisely formulated.) Still, Sidgwick suggests that some hints about the origins of these three principles were foreshadowed in his preceding arguments in *ME*. We therefore need to clarify the relationship between such preceding arguments and these three principles. We also need to clarify the differences and similarities of these three principles. On the one hand, Sidgwick points out that these three principles share a common element. On the other hand, however, one might naturally assume that these three principles state different things because Sidgwick presents them as separate principles. In addition, Sidgwick elaborately argues that the Principle of Rational Benevolence can consist of two truths that are intuitively apprehended, whereas the Principles of Justice and Rational Self-Love are straightforwardly apprehended by philosophical intuition. Keeping these points in mind, we will now examine each fundamental principle and what kinds of arguments are used to develop it.

6.1 Description and interpretation of the principles

The three principles are stated in several formulations in *ME* Book III, Chapter 13, and also foreshadowed and restated in other places in *ME*. Those formulations and explanations are neatly summarized in a list by Schneewind, who wrote a well-known commentary on Sidgwick's ethics (Schneewind 1977, pp. 295–7. See also Shionoya 1984, pp. 155–9). In order to understand the meaning of each principle, however, it is not sufficient for us just to look at Schneewind's list, which even includes the expressions that Sidgwick only elliptically adopted. Actually, hints to understand the derivation and the real intention of the three principles are to be found in Sidgwick's analyses of 'right', 'ought', and 'good', which were previously explained. Schneewind, however, does not make such an in-depth investigation of implicit relationships among different passages in *ME*. I will attempt to clarify the precise meaning of the three principles, including their relationship to the previous conceptual

analyses and concentrating only on those formulations which Sidgwick himself admitted to be precise.

6.1.1 The Principle of Justice

The Principle of Justice can be derived by clarifying our intuition that was already manifested in our analyses of the concepts of 'right' and 'ought'. This principle states that individuals ought to be treated equally – equally, in terms of their *logical* treatment – in making right- or ought-judgments.

In his analyses of 'right-' 'ought-' judgments in Book I, Chapter 3 of *ME*, Sidgwick suggested 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' (*ME* p. 33). There he pointed out that, even when made by a particular person regarding a particular action, an ethical judgment is to be similarly judged by other persons unless some error is found, and to be equally applicable to any other similar action (*ME* pp. 33–4. See also 4.2.3 of the present book). This point evolves into 'one practical rule of some value, to be obtained by merely reflecting on the general notion of rightness, as commonly conceived' in *ME* Book III, Chapter 1, Section 3 (p. 208). It is important to understand exactly what Sidgwick claims here. What Sidgwick insists is *not* that one must always judge the same action to be right for other persons once one judges it to be right for oneself. It is rather that one must make the same judgments in both situations *unless one presents a reasonable explanation* for judging them differently. If there is some important difference in nature or circumstance between the two situations, one may make quite opposite ethical judgments about what ought to be done by or to a person in each instance. For example, if an elderly person with no relatives wishes to live in a special nursing home while another aged person, who has a family who can take care of him, wishes to spend his remaining days at his own house, it is quite appropriate to judge that the first person ought to live at the nursing home but the second ought not to do so. It is not cogent, however, to make different ought-judgments when there is no such recognizable difference in nature or circumstance. If there is another elderly person who, like the second person, has family and wishes to live at his own house, and if there are no special differences between the second and the third persons, it is not proper to insist that the third elderly person ought to go to a nursing home even though the second ought not to do so. Thus, Sidgwick claims:

> We cannot judge an action to be right for *A* and wrong for *B*, unless we can find in the natures or circumstances of the two some difference

> which we can regard as a reasonable ground for difference in their duties. If therefore I judge any action to be right for myself, I implicitly judge it to be right for any other person whose nature and circumstances do not differ from my own in some important respects.
>
> (*ME* p. 209)

Sidgwick does not clearly state what this 'reasonable ground' should be. Judging from Sidgwick's explanation of 'reason', however, we can construe it as a kind of difference that truly reflective people would commonly recognize. Such a difference is not what a single individual can arbitrarily determine, but one in which every reflective person would understand as a reasonable ground for assigning different kinds of treatment.

A more precise formulation of the Principle of Justice is presented in Book III, Chapter 13 in *ME*. Here Sidgwick restates the principle given in Book III, Chapter 1, Section 3, as follows:

> One such principle was given in chap. i. § 3 of this Book; where I pointed out that whatever action any of us judges to be right for himself, he implicitly judges to be right for all similar persons in similar circumstances. Or, as we may otherwise put it, 'if a kind of conduct that is 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'.
>
> (*ME* p. 379)

We should note that Sidgwick presents this principle in two ways here. He also suggests that this principle applies not only to what ought to be done *by* individuals, but also to what ought to be done *to* individuals. To integrate this point with the above explanation, the principle can be reworded as follows:

1. If I judge an action done by or to myself to be right for me, I implicitly judge it to be right for all similar persons in similar circumstances.
2. If I judge an action done by or to myself to be right for me but wrong for someone else, that must be because there is some difference in the natures or circumstances between the two, other than the fact that they are two different persons.

Proposition (1) states that a particular 'ought-' or 'right-' judgment implies a universal judgment that is applicable not only to a particular

individual to whom that particular judgment refers, but also to *all* similar individuals placed in similar situations. Proposition (2) is a contra-positive of (1), stating that, when two 'ought-' judgments differ, the situations or persons referred to in those judgments should prove to be somewhat *dissimilar*. Whether expressed in the first or the second proposition, the Principle of Justice states the requirement that similar people should be equally *logically treated* in ought- or right-judgments. The term 'Justice' denotes this fairness in making ethical judgments. This fairness, however, is not related to the evaluation of people's good, but only to the *logical* treatment of the individuals referred to in the judgments. That is, we have to do justice to two individuals, by making the same ought- or right-judgments about them when there is no notable difference in the nature or circumstance of the two persons and their actions, except the fact that they are two different individuals.[1]

According to Sidgwick, just by using proposition (1), we can test the ethical judgment that we are about to make. When we ask ourselves if we are ready to judge that the same act ought to be done by similar individuals in similar situations, we often notice that our ought- or right-judgments are but a result of our own biases, and that we do not seriously wish to make such judgments that have universal application (*ME* p. 209). In reality, however, two actual situations cannot be strictly the same; they always have some differences between them. Sidgwick is aware of this fact, but believes we can still suggest that, if we are to make different ought-judgments for two different individuals, we must be able to point out some 'dissimilar' points between the two, *other than the mere fact that they are different persons*. When this principle is applied to a situation where only two people are involved, it implies the following (*ME* p. 380):

3. 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.

This formula indicates that, when we cannot state any difference in situations except that A and B are two different individuals, we have to make the same ethical judgments in similar situations *in which the positions of the two individuals are exchanged*.

According to Sidgwick, this Principle of Justice merely requires that one shoulder a certain *onus probandi* in applying to another such a treatment

that one would complain about if applied to oneself (*ME* p. 380). Sidgwick insists, however, that commonsense has sufficiently recognized the practical importance of this maxim, and that it seems self-evident to him that this maxim is true as far as it is stated as shown above.

However practical it is, the Principle of Justice only states a formal requirement to be observed when one makes a judgment that a certain action ought to be done, *after* one has chosen that action among possible alternatives. This principle itself does not offer us a guide for deciding what action to choose. In order to get such guidance, we still have to look for other fundamental principles. Next, Sidgwick introduces the Principle of Rational Self-Love.

6.1.2 The Principle of Rational Self-Love or Prudence

The maxim of Rational Self-Love clarifies the intuition that we hold when we state one ought to seek the good on the whole for oneself.

In 4.3.6 of this book we suggested that 'the good on the whole for oneself' is what one would prefer when one only considers one's own existence and when one's desire is in harmony with reason. We then stipulated it as what one would now desire if one considers the good of oneself in the future as well as one's present good. However, I suggested there that this definition does not indicate *how* such future good should be weighed – it does not necessarily imply that one must put *greater or lesser weight* on one's future good. The Maxim of Self-Love requires us to attach *equal weight* to one's present and future good. To put it more precisely, if we judge a good at a certain point of time to be as great in size as that at a different point of time, the Principle of Rational Self-Love requires us to treat them as having equal weight. We have already seen that 'good' is a quantitative notion that can be expressed as having a 'greater' or 'lesser' degree. The Principle of Self-Love dictates that we should evaluate our good at one time or another time *according to their quantities*.

We most often put this maxim in the proposition 'that one ought to aim at one's own good' (*ME* p. 381). However, if this proposition is simply interpreted as claiming that one ought to aim at one's good at present, assuming that one only considers one's present self, it becomes a mere tautological proposition which is true by the definition of 'one's good at present', as analyzed in 4.3 of this book. If so interpreted, we cannot regard this as a significant practical principle. Again, we can interpret the same proposition as claiming that one ought to aim for what *one now judges* to be the greatest attainable good on the whole; but this is also tautological as we ascertained when we defined 'the good on

the whole', and therefore it cannot be regarded as a significant proposition. Nevertheless, we implicitly assume certain restraints when we estimate the amount of certain good on the whole. The maxim of Rational Self-Love suggests one such restraint.

When I talk about my good on the whole, I recognize that a good for my *future* self might be different from that for me *at present*, and then my reason imposes certain restraints on my present judgment as to how I make this evaluation. In doing so, I suggest that I at present ought not merely to *foresee* my future, but also to *give weight* to my future good. I thereby imply a certain nontautological proposition, which provides a substantial dictate that I not put biased weight on my present. This principle of 'impartial concern for all parts of our conscious life' (*ME* p. 381) offers practical guidance for us, who often fail to be prudent. Sidgwick believes that we would all admit the validity of this principle, though we do not always follow it. This is the intuition that makes up the maxim of Rational Self-Love.

This intuition is stated in several ways, as shown below (see *ME* pp. 381, 383):

1. Hereafter [= Good in the future] *as such* is to be regarded neither less nor more than Now [= the good at present].
2. The mere difference of priority and posteriority in time is not a reasonable ground for having more regard to the consciousness of one moment to that of another.
3. That a smaller present good is not to be preferred to a greater future good (allowing for difference of certainty).

Sidgwick sometimes uses terms such as 'consciousness' or 'conscious life' as seen in formula (2). This can be understood as a shortened expression of 'the different "goods" that succeed one another in the series of our conscious states' (*ME* p. 382) or a part of such good at a certain point of time. We can understand the reason why Sidgwick sometimes uses the term consciousness instead of good if we recall the distinction between 'the good for me at present' and 'my good on the whole'. The present and the future good mentioned in formula (3) means 'the good for me at present = what I would desire when I only consider my present self without thinking about ulterior consequences' and 'the good for my future self = what my future self would desire when I consider myself at that time', respectively. 'The consciousness of one moment' in formula (2) is an abbreviation of 'the good for me at present' that I at that moment would be conscious of. What I judge

to be *good for me at one moment* is quite different from what I judge to be *my good on the whole*. Probably for that reason Sidgwick often prefers to call the former 'the consciousness at that moment' instead of simply calling it good.

In any case, the crucial point of the Principle of Rational Self-Love, which is expressed by the formulas shown above, is that the scale to *weigh* different goods ought to be *impartial over time* in our measurement of an individual's good on the whole. We have already seen, in the previous analysis, that the proposition that 'I ought to prefer the greater rather than the lesser good for myself at present' is true by the definition of good, but there I also pointed out that Sidgwick did not claim, at that point, anything about how to deal with 'the good for one's future self' in relation to 'the good for one's present self'. The Principle of Rational Self-Love prescribes that we should not shift the weight on two different goods *just because they occur at different times*, if they have the same quantity and degree of certainty. This presupposes that we can more or less estimate the size or the quantity of goods at different points in time. This Principle of Rational Self-Love contains quite a different type of requirement from that of the Principle of Justice, that is, the requirement to measure the *quantities* of good. It should be noted that 'good' has no substantial content at this stage, except that Sidgwick defined good as what is desirable. He has neither specified how we can determine the size or quantity of different goods at different times.[2] What the Principle of Rational Self-Love states is that, as far as we admit that two goods are of the same size, we must not attach more importance to one and less to the other just because the latter occurs in the future. We should also note that, as several writers have pointed out (see, for example, Seth's and Hayward's claims explained in Schneewind 1977, Ch. 10 pp. 304–5; Shionoya 1984, p. 157, etc.), this principle itself does not state that we ought to *maximize* the total goods at all points of time. The Principle of Rational Self-Love necessitates impartiality in weighing goods at different points of time, but does not dictate the *maximization* of those goods.

This principle is referred to as *Self-Love* or *Prudence*, for Sidgwick derived it from cases in which a person considers his or her own good on the whole. However, the essence of this principle is the equal weighing of the same size of goods throughout time. In other words, one 'ought not to prefer a present lesser good to a future greater good' (*ME* p. 383). This maxim can also be applied to cases in which I judge someone else's good on the whole throughout his or her life. As a general explanation of the Principle of Rational Self-Love, Sidgwick states that

this is an intuitive principle suggested when we talk about '"good on the whole" – of *any* individual human being' (*ME* pp. 381–2. Emphasis added); he does not claim that this 'individual human being' should be restricted to 'me' or 'oneself', and exclude any *other* individual human being. In fact, if I am to evaluate someone else's good on the whole, it would be odd for me to prefer his present lesser good to his future greater good just because they occur at different times.

6.1.3 The Principle of Rational Benevolence

The maxim of Rational Benevolence is derived from the following two self-evident intuitions, (1) and (2), and its precise formulation is expressed by (3) (*ME* Bk. 3 Ch. 13 Sec. 3 p. 382).

1. 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.
2. 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.
3. Each one is morally bound to regard the good of any other individual as much as his own, except in so far as he judges it to be less, when impartially viewed, or less certainly knowable or attainable by him.

In one passage Sidgwick calls proposition (1) a self-evident *principle*. If we may regard (2) as a kind of principle as well, we can say that (3) is a practical maxim that resulted from a combination of two intuitive principles. However, (1) is an explanatory principle about the relationship between one's good and another person's good. In contrast, (3) is a practical principle that provides a certain guide for action, and for this reason the third formulation is often called a *maxim*, that is, the maxim of Rational Benevolence.

This maxim represents the requirement that one ought to impartially evaluate a good for oneself and a good for someone else. More precisely, it requires that the scale to *weigh* different goods ought to be *impartial for various individuals*. Its main idea is briefly expressed in the statement that 'I ought not to prefer my own lesser good to the greater good of another' (*ME* p. 383). Here, we should remember that 'the good of each individual' does not directly correspond to the satisfaction of each person's *actual* desire. The Principle of Benevolence does not claim that one ought to impartially weigh the object or the satisfaction of personal desires. What

is considered here is the 'good' for each individual, namely, what he *would* desire if he considers only himself, and if a certain ideal condition is met, that is, if he fully recognizes all feasible alternatives and the desire-satisfaction resulting from them. That is to say, a good for each individual is the *rational* object of his *well-considered* desire.

Here again, we should note that 'good' is given no substantial content at this stage, nor is any explanation given of how we can determine whose good is greater or lesser than someone else's good. The essence of the Principle of Rational Benevolence is that, if one admits that one's good and another's good are of the same size, one should never give lesser weight to the latter good just because it is another's. What this principle requires is impartiality in weighing different 'goods' of different people. This principle itself does not claim that one ought to maximize the good for all people.

One more thing to note about the Principle of Rational Benevolence: we admit that we should put impartial weight on another's good as well as on our own good *only when* we go beyond our own good and take an overall impartial viewpoint (what Sidgwick calls 'the point of view of the Universe'). We would all admit that, *once we take this viewpoint*, we have to evaluate people's good according to the Principle of Rational Benevolence. However, whether we actually accept this viewpoint or not is a different question. This opens an escape route for egoism, which will be explained later in this book (7.5).

6.2 Distinctive features of each principle

According to Sidgwick, one salient element common to these three principles is 'the relation of the integrant parts to the whole and to each other' (*ME* p. 382). In the Principle of Justice, he posits a 'Logical Whole' that is comprised of individuals, who are equally treated in the logic of ethical judgments (*ME* p. 380). The Principle of Rational Self-Love deals with the 'good on the whole' of a single individual, which is a serial awareness that consists of different 'goods' succeeding one another (*ME* pp. 381–2). The Principle of Rational Benevolence is concerned with the 'good on the whole' that is made up of 'goods' of all individual human or sentient beings (*ME* p. 382). Thus, all three principles are said to represent the relationship that individuals or their pursuits have as parts to their wholes and the relationship among the various parts (*ME* pp. 382–3).

Nevertheless, there are some critical differences among these three principles, especially between the Principle of Justice and that of Self-Love or

Benevolence. For this reason, Sidgwick had to present not one but three principles separately – unlike R. M. Hare, who proposed a single principle which he called the universalizability of evaluative judgments. To be noted here is the difference between a 'logical whole' and a 'mathematical or quantitative whole', which will be explained below. As previously stated, the Principle of Justice is obtained by considering the similarity of particular individuals that make up a logical whole or genus. In contrast, Sidgwick claims that the Principles of Self-Love and Benevolence are derived by considering the similarity of the elements of mathematical or quantitative wholes (*ME* p. 381).

The meaning and the importance of this difference were scarcely emphasized in the previous studies of Sidgwick's ethics. I would point out that this difference was even erroneously interpreted by some researchers. For example, Professor Shionoya, one of the leading researchers of Sidgwick in Japan, is apparently wrong in claiming that 'the three axioms share the logic of a mathematical whole in common', for he misses the point that the mathematical or quantitative whole relates only to the Principles of Self-Love and Benevolence, while the Principle of Justice is concerned with a *logical* whole (Shionoya 1984, pp. 156–7, 159. See also note 3 in this chapter). In a passage where he has just finished the explanation of the Principle of Justice and goes on to explicate the Principles of Self-Love and Benevolence, Sidgwick clearly states as follows:

> The principle just discussed, which seems to be more or less clearly implied in the common notion of 'fairness' or 'equity,' is obtained by considering the similarity of the individuals that make up a Logical Whole or Genus. There are others, no less important, which emerge in the consideration of the similar parts of a Mathematical or Quantitative Whole.
>
> (*ME* Bk. 3 Ch. 13 Sec. 3 pp. 380–1)

This pairing of a 'logical whole' and a 'mathematical or quantitative whole' seldom appears in other places in *ME* (the phrase 'a mathematical whole' is used once again in the fifth line from the bottom of *ME* page 381). Therefore, it is not surprising that this difference has been unnoticeable. Nonetheless it is very clear from the passage cited above that the mathematical whole is not what is common to the three principles but what relates only to the Principles of Self-Love and Benevolence, and that the Principle of Justice deals with a separate notion of a logical whole. In fact, there is a huge, unrecoverable gap between the two kinds

of wholes and between the Principle of Justice and the other two principles that correspond to these wholes. Contrary to Professor Shionoya's claim that the Principles of Self-Love and Benevolence can be obtained by *applying* the Principle of Justice to the notion of good (Shionoya 1984, pp. 156–7), the Principles of Self-Love and Benevolence have one crucial element that can never be procured by merely 'applying' the Principle of Justice. I should acknowledge that this point was first suggested to me by Uchii in 1997.[3] Still, I would like to stress this point as 'Independent Interpretation' in this book, 'Independent' denoting the conceptual independence of the logical whole from the mathematical ones, and of the Principle of Justice from the other two principles. This will later play a critical role in my arguments against some claims of contemporary utilitarianism.

My introductory remark was lengthy, but the point is simple. What kind of difference among the three principles is suggested by Sidgwick's claim of logical and mathematical wholes? To this question Independent Interpretation gives the following answer.

The Principle of Justice concerns the logical property of judgments about right action. An ought- or right-judgment about a particular individual person implicitly applies to *anyone* who belongs to a certain definable class that is made up of individuals similar to him. A logical whole is the class of similar individuals to whom an ought- or right-judgment equally applies. A part of such a whole is each individual to whom the ethical judgment is applied. The Principle of Justice deals with the logical treatment of individuals, or with the question of whether an ethical judgment applies to a certain individual or not. Here the notion of quantity is utterly irrelevant, since the question addressed by the Principle of Justice is not a matter of *degree*, but a matter of *whether* a judgment applies to an individual *or not*.

In contrast, the Principles of Self-Love and Benevolence deal with the question of how to treat the *quantities* of good; the former principle is concerned with the quantity of 'my good at each moment', which constitutes my good on the whole, and the latter with that of 'the good of each individual', which sums up the good for all individuals. Thus, Sidgwick views these wholes not as *logical* but as *mathematical* or *quantitative* wholes. Whereas the application of right-judgments is not a matter of degree or quantity, 'good' is basically a quantitative notion that requires a scale to measure its size and to compare it with others. The Principles of Self-Love and Benevolence state that one and the same scale ought to be equally applied to any point of time and to any individual. By stating this, these principles stipulate *how* one ought

to pursue good in deciding what one ought to do in a situation where people's present and future 'goods' are in question.

To summarize the differences among the three principles: the Principle of Justice represents the requirement that, in making ethical judgments using terms such as 'right' or 'ought', individuals ought to be equally logically treated. The Principle of Rational Self-Love states that, in considering an individual's good on the whole, the scale to measure different goods ought to be impartial *over time*. Sidgwick introduces here a requirement for measuring the quantity of good at different times. The Principle of Rational Benevolence states the additional requirement that, in considering people's overall good, the scale to measure goods ought to be impartial *for different individuals*.

One may perhaps think that these differences are, once clarified, simple and trivial. In Chapter 8 of this book, I shall reveal the crucial significance of Independent Interpretation, and highlight the acuteness of Sidgwick's insights in presenting these three principles through my reexamination of Hare's moral philosophy.

6.3 The three principles and the five conditions

Sidgwick explains the three principles as 'truly self-evident and significant' truths. For him, these principles satisfy the five conditions described in the previous chapter, which are necessary for propositions to be self-evident and significant. Schneewind has already carried out this test in a somewhat brief manner (Schneewind 1977, Ch. 10, p. 297), yet I intend to retest these principles in the light of these five conditions, based on my own understanding.

First of all, for a proposition to be truly self-evident, every term used in that proposition should be clear and definite. As for the Principle of Justice, one key term used is 'right'. Sidgwick has already analyzed the concept of right (see Chapter 4, Section 2 of the present book), and, though this concept turned out to be indefinable, he adequately clarified its logical properties manifested in its various uses. Other terms, such as 'action', 'persons', 'nature', and 'circumstances', are also *either* analyzed in the preceding chapters of *ME* or appear to have definite meanings in the context of this principle, and most people will not question their meaning. One may perhaps think that the term 'similar' is ambiguous and unclear, but what the Principle of Justice claims is that, *if a person judges two situations to be similar*, he must make the same judgments on them. Whether they are similar or not is determined by the person who makes the judgment, so they cannot be interpreted by other individuals in different ways.

As for the maxim of Rational Self-Love, its key terms 'good' and 'one's good on the whole' have already been analyzed. Temporal notions such as 'future', 'present', and 'time' also play important roles in this principle, but most people would not disagree on their common meaning. Some may think that the expressions 'consciousness' and 'conscious life', which are used in some passages to restate this principle, need to be explained; still, judging from Sidgwick's own explanation of this principle, it is clear that these expressions denote a certain mental state in which one has preferences toward certain good at that moment.

The maxim of Rational Benevolence focuses on the 'good on the whole', which means people's overall good as already defined. The expression 'greater (or smaller) good' derives from the fact that good is essentially a comparative notion, and it means the good that ought to be preferred and pursued *more* than others. 'A rational being' is, as we have seen, such a being that recognizes truth and is motivated based on that recognition. His somewhat puzzling phrase, 'the point of the view of the Universe', has a touch of rhetoric. It is followed by a parenthesized proviso 'if I may say so', and it actually means an impartial viewpoint in which one temporally puts aside one's personal likes and dislikes and considers the preferences of all individuals. What one is 'morally bound' to do is what one ought to do when one takes such an impartial point of view. Thus, the terms that are used to describe these three principles are all carefully defined, clarified, or commonly understood.

Second, the fundamental moral principles should not be tautological, but have substantial content that can serve as a guide to determine what one ought to do. The Principle of Justice is not a tautological proposition which states it is right to do the right thing. Rather, this principle conveys substantial information about what is *required* when one intends to make an ought- or right-judgment, even though at first glance this principle seems to simply describe the basic use of the indefinable notion 'ought' or 'right'. This requirement provides a binding dictate for a person who is already using these terms, by asking the question of what one ought to do. The Principles of Self-Love and Benevolence are also not tautological propositions which state that one ought to aim for what one now judges to be the greatest good on the whole. Rather, they give us certain binding dictates about how one ought to measure the good on the whole. In judging good on the whole, we ought to give equal weight to our present and future good, and to the good for other individuals. This is not what we do naturally, but what we must will to do. These principles therefore give us nontautological but substantial dictates.

Third, we must ask whether these principles remain self-evident upon careful reflection and fourth, whether there is a universal or general consensus about their validity. For example, can these principles be accepted even if we set aside our conventional, legal, and traditional beliefs? Sidgwick claims that, while most moral principles differ from culture to culture and from period to period, these three principles seem to hold in any society at any time. Furthermore, he ascertains that the same principles of Justice and Benevolence can be found in the writings of Clarke and Kant, and finds the Principle of Self-Love in the work of Butler, who claimed that prudence is also a manifest duty.

However, we have to clarify the exact meaning of 'any reflective person will admit these three principles'. Everyone would admit that (1) one ought to abide by the requirement of the Principle of Justice *if* one is to make a 'right' or 'ought' judgment, that (2) one ought to follow the Principle of Self-Love *if* one considers an individual's good on the whole, and that (3) one ought to comply with the Principle of Benevolence *if* one considers people's overall good from an impartial viewpoint. Sidgwick claims that *anybody* accepts these three principles, whether or not he/she is an egoist or a deontologist. This simply means that *anyone* would admit the former part of each proposition *as far as the latter condition holds*.

So, our test has been quite successful so far. Finally, what about the condition that the propositions accepted as self-evident should be mutually consistent? It should be first noted that these three principles do not logically contradict each other, for they deal with quite different dictates that operate on different dimensions and under different conditions. The Principle of Justice concerns the logical property of moral judgment, and the other two deal with the way to pursue good. The Principle of Self-Love presupposes that one would take the viewpoint of one's good on the whole, and the Principle of Benevolence assumes that one would take the viewpoint of people's good on the whole. In general, the proposition that one ought to do A under condition B cannot logically contradict the proposition that one ought to do C under condition D.

However, Sidgwick later admits that there is a possibility of conflict between these propositions when applied to practice. He admits that a case may occur in which an action dictated by the Principle of Self-Love runs counter to one dictated by the Principle of Benevolence. This conflict develops into a problem called the 'dualism of practical reason' in the final chapter of *ME*, but Sidgwick conceals the possibility of such practical conflicts when he introduces these three principles. So, for the

time being, let us provisionally assume that, unlike most general moral principles obtained by our dogmatic intuition, the three principles that were obtained by our philosophical intuition appear undoubtedly self-evident and significant. After all, these three are at the core of our reflective commonsense, whatever method of ethics we usually use and whatever general moral rule we normally follow.

6.4 The three principles and the three methods of ethics

Sidgwick explains the relationship between the above three principles and the three methods of ethics, as follows:

> The axiom of Prudence, as I have given it, is a self-evident principle, implied in Rational Egoism as commonly accepted. Again, the axiom of Justice or Equity as above stated –'that similar cases ought to be treated similarly'– belongs in all its applications to Utilitarianism as much as to any system commonly called Intuitional: while the axiom of Rational Benevolence is, in my view, required as a rational basis for the Utilitarian system.
>
> (*ME* Bk. 3 Ch. 13 Sec. 5 pp. 386–7)

The three fundamental principles are what everyone, whatever method of ethics he takes, would accept as truth. In particular, the Principle of Justice is a logical rule with which anyone must comply whenever he is making a judgment about what 'ought' to be done or what it is 'right' to do, whether he adopts the method of dogmatic intuitionism, egoism, or utilitarianism. The Principles of Self-Love and Benevolence are supposed to be accepted by everyone, given the conditional clauses '*if* one considers an individual's good on the whole' and '*if* one considers people's overall good'.

The method of egoism especially calls for the Principle of Rational Self-Love. Needless to say, the view that one ought to seek one's lifelong happiness must require that one give equal weight to one's future good and one's present good, having interpreted good as one's own happiness or pleasure.

However, the Principle of Self-Love is by no means exclusive to egoism. This principle is presupposed not only in egoistic but also in utilitarian methods, for it is unacceptable even for utilitarianism that one put disproportionate weight on an individual's good at a particular time *just because* it occurs *then* and not at other times. What makes utilitarianism distinct from egoism, however, is the Principle of Rational

Benevolence. We cannot explain the utilitarian ideal of equal consideration of people's happiness without presupposing the requirement that one ought to give equal weight to others' good (which is interpreted as their happiness or pleasure in utilitarianism) as well as to one's own good, in proportion to its quantity. Thus, we have hereby obtained one basis for utilitarianism.

7
Philosophical Foundations of Utilitarianism

We have finally come to the point of elucidating the basic structure of utilitarianism by making the most of Sidgwick's analyses and arguments. In *The Methods of Ethics*, utilitarianism gets its theoretical foundations through philosophical inquiry and gains external support from a careful examination of common-sense morality. In this chapter we will examine both ways of verifying utilitarianism. We will see that the previously described conceptual analyses and the three intuitive fundamental principles, *plus* the proof of hedonism (to be explained later in this chapter), construct the essential components of utilitariaism, namely *consequentialism* and *the principle of maximizing the sum total of people's pleasure*.

7.1 Consequentialism and the maximization principle

7.1.1 Consequentialism

Utilitarianism is a kind of consequentialism, in that it judges the rightness of an action by evaluating the goodness or badness of its *ulterior* consequences (i.e., other results than the performance of an action). As far as our ordinary judgments go, however, the right actions are not always those which pursue *ends* to be realized by those actions or their consequences, for we sometimes consider actions that unconditionally comply with certain duties to be the right actions; in such cases the ulterior consequences or the 'ends' of actions do not come to mind. However, Sidgwick claims that, on reflection, we will notice that we need to refer to the notion of 'good', or the ends that our actions ought to seek, in order to decide in a systematic and consistent manner what we really ought to do.

The crucial arguments on this point can be found in Book III of *ME*, where Sidgwick reveals that the rules of common-sense morality that

are commonly regarded as unconditional duties cannot become truly significant self-evident principles, no matter how much we refine them. In Book III, Chapters 3 through 10 of *ME*, Sidgwick examines particular duties and virtues of common-sense morality. His list of such duties and virtues is quite exhaustive, and it includes wisdom, benevolence, special duties to particular people (such as parents, children or friends), justice, observance of laws or promises, truthfulness, generosity, tolerance, temperance, purity, courage, humility, and so on. Sidgwick always tries to precisely define each one of these duties and virtues, and asks if we can create a coherent moral system only by clearly formulating these common rules, as dogmatic intuitionism claims.

As a result of such examination, however, it turns out that all these duties and virtues do not satisfy the five conditions for self-evident and significant axioms, in so far as we maintain that they are always valid apart from their ulterior consequences. That is, on reflection, we find that:

1. The terms and concepts used to express duties or virtues are often unclear and ambiguous. For example, even if we admit that we have a special duty to be kind to our friends, we cannot clearly state which people are covered by the term 'friends'. It is also unclear whether 'a promise' is valid only when a promisor and a promisee understand it in exactly the same sense, or whether it is valid even when there is some misunderstanding between them. We cannot clearly determine whether 'telling a lie' only means making an utterance which explicitly contradicts facts, or whether it also includes the implicit suggestion of a certain erroneous fact.
2. These common moral rules often conflict with other common moral rules, as in conflicts between the duty to keep a promise and that to help others, or between equal distribution and distribution by merit in distributive justice. In such cases of conflict, we cannot arbitrate between them by appealing to those common moral rules themselves, without referring to the ulterior consequences of our actions.
3. Furthermore, every rule has certain limitations or exceptions. It is often said that we do not have to keep our promise when a promisee voids it, when that promise was made by fraud or coercion, or when our circumstances have significantly changed, etc. However, we often disagree on the proper extent of such limitations and exceptions, and it is extremely difficult to define them clearly. When we start to carefully reflect on these rules to determine their limits and exceptions, however, we come to realize either that we can no longer

regard those rules as self-evident, or that there is no general consensus about such limitations and exceptions.

These arguments imply that we cannot establish a systematic ethical theory by means of such a nonconsequentialist approach, which regards common moral rules as absolutely valid in themselves. Thus, when we analyze common duties and virtues, we realize the need for some fundamental principle *other than* the rules of such duties and virtues, in order for us to determine the right actions in a consistent and systematic way.

However, from this examination Sidgwick does not immediately conclude that consequentialism is the right answer. The moral judgments that we usually make include not only the nonconsequentialist one that an action is right regardless of its consequences because it is one's duty or virtue but also another type of nonconsequentialist judgment that an action is right regardless of its consequences when the agent's motive is good. In 1.2.4 of the present book I stated that one's motives were not our primary concern when we decide the right actions. Now, however, it has become clear that we cannot precisely decide what we ought to do by abiding by common duties or virtues, and we may well conclude that we cannot strictly judge the rightness or wrongness of actions, and return to the view that we can only judge the goodness or wrongness of people's motives. Sidgwick examines this possibility in Book III, Chapter 12 in *ME*, but he finally dismisses it. This is because, for one thing, at least some motives cannot be judged to be good or bad without considering what actions and consequences they will bring about. For example, we will not be able to determine whether a person's motive to do justice is good if we do not know what it means to do justice and what results this motive actually brings about. For another thing, those who attach great importance to an agent's motives quite often classify various types of such motives, from higher to lower, and thereby attempt to decide the goodness and wrongness of a person's motive in a particular situation. However, the rankings of motives are different from person to person, and no agreement can be expected. Some rank a benevolent motive as the highest, but there are others who, like Kant, regard the motive of pure respect for duty as supreme. Sidgwick illustrates these difficulties by citing the list of motives by his contemporary, Martineau, who claims that motives range from the lowest to the highest as they move from love of ease and sensual pleasure, through appetites, fear/resentment, love of power, generosity, and compassion, and so on, to the sentiment of reverence. We do not seem, however, to make all our moral judgments in accordance with his chart. When we condemn a person

who overeats, we do so not because the motive to eat is lower than other motives but because we know from experience that overeating tends to threaten one's health. We might blame a captain who did not reduce a ship's speed, for laziness (or love of ease), when a ship was wrapped with a dense fog. We blame him not because his love of ease is lower than his fear of an accident but because the accident that could result from not slowing down is very serious (*ME* pp. 369–70). Furthermore, when Martineau's lower and higher motives conflict with each other in a particular situation, we do not always think the higher one should prevail. For example, though Martineau ranks resentment as lower than compassion, we may judge a person's resentment against an injustice to be a more appropriate motivation than his compassion. Such a judgment can be made by closely examining the particular situation in which a conflict between motives is taking place. If such a serious dilemma occurs, it is unlikely that we can judge the morality of an action just by comparing different motives. A conclusive comparison is not the one between the motives driving a situation but the one between alternative actions and their consequences.

Thus, after all we have to deal with the rightness and wrongness of action. It has turned out, however, that we cannot determine it in a systematic manner just by appealing to common duties and virtues that we usually believe should be observed unconditionally. Therefore, if we are to discern truly right actions, we have to find higher principles other than these common rules of duties and virtues. With this awareness of the ultimate fragileness of common moral rules, we have reached, using our philosophical intuition, the three practical principles, which are the Principles of Justice, Rational Self-Love and Benevolence. Of them the Principles of Rational Self-Love and Benevolence refer to certain ends, or 'goods', which denote something different from a mere observance of duty, and require us to treat different 'goods' equally within a certain dimension. If these are the principles that we should consider when deciding what we ought to do, it follows that we should consider the various 'goods' that our actions would bring about whenever we attempt to decide what we ought to do. This idea is what we now call consequentialism. The right actions are right because they bring about certain good consequences. Thus, after rejecting two nonconsequentialist ideas, that is, dogmatic intuitionism and the motive theory, we now have a rational basis to argue for a consequentialist ethical position – or a position that considers consequences in some way – based on the fundamental ethical principles that we have found through our philosophical intuition.

7.1.2 The maximization principle

However, a position to consider consequences does not simply mean that it is a utilitarian position. *Any* position that *more or less* considers an action's consequences might be called consequentialism, but utilitarianism is a view that seeks to *maximize* people's good on the whole that will result from an action. How can this 'maximization principle of people's good' be theoretically derived?

Sidgwick himself seldom discusses this question of how the utilitarian maximization principle was derived. Assuming our previous interpretations and analyses, however, I think we can reconstruct the argument to support this principle in a coherent way.

The previously described Principles of Justice, Rational Self-Love and Rational Benevolence are all concerned with deciding the right action. These three principles, however, are but principles that require some kind of impartiality. The Principle of Justice does not state anything about good; it simply requires us to treat individuals equally in the logic of applying ought- or right-judgments. The two principles that are supposed to deal with good, that is, those of Self-Love and Benevolence, only state that one should impartially treat different 'goods' within a certain sphere, and do not dictate that one must seek to *maximize* such good on the whole.

The conclusion that the right action is the one that maximizes people's good can be derived if we recall the meaning of 'the greatest good' (4.3.2 of the present book) as well as the three self-evident axioms. The idea of maximization is derived from the concept of good. By the very definition of 'the greatest good', one ought to perform an act that one now judges to bring about the greatest attainable good, in so far as one is going to choose among alternative actions. In other words, one ought to maximize, as much as possible, what one now judges to be the good on the whole. I suggested earlier that this is a tautological truth. Utilitarianism does not, however, simply require us to maximize good but to maximize *people's* good. This distinctive feature of utilitarianism, that one ought to maximize good for *people including oneself and others*, is not tautological, unlike the proposition that one ought to maximize what one now judges to be the good on the whole. This utilitarian proposition holds when the Principles of Rational Self-Love and Benevolence lay certain restraints on *how* 'the good on the whole' should be maximized.

The Principles of Rational Self-Love and Benevolence require that one ought to give impartial weight to individual 'goods' in proportion to their size, regardless of when they happen and to whom they apply. When we accept this requirement, it is natural for us to judge 'the good

on the whole' to be the sum total of individual 'goods' (more precisely, 'goods' of all the different individuals and at all the different times) that are impartially treated in proportion to their size, or so Sidgwick believes. Here Sidgwick actually relies on two basic assumptions that (1) 'goods' are something that can be added together, and that (2) a whole is the sum total of its parts. One might object that these two assumptions are not necessarily self-evident – I shall deal with this point in 10.3 of this book.[1] For the sake of argument, however, let us at least temporarily accept Sidgwick's idea that 'the good on the whole' is something that somehow aggregates individual goods, and that such an aggregation is supposed to be *the sum total*, as the simplest and the easiest way of thinking about a whole. When we combine this claim that 'the good on the whole is the *sum total* of individual goods' with the philosophical intuition that 'from an impartial viewpoint, such individual goods should be equally treated in proportion to their size' *and* the previous argument that 'one ought to *maximize* the good on the whole', we can obtain the utilitarian maximization principle, which asserts that one ought to *maximize*, as much as possible, *the sum total* of individual goods that are *impartially measured*.

This is how we can establish the basic structure of utilitarianism, or the maximization principle of the sum total of people's good on the whole. However, the utilitarian ethical theory that Sidgwick supports is a *hedonistic* version of utilitarianism, which pursues people's pleasure as people's good. We have to further examine Sidgwick's argument that we should ultimately aim at pleasure, and pleasure only, as our ultimate good. So in the next section I will explicate Sidgwick's hedonism and his 'proof' of it.

7.2 Hedonism

7.2.1 What pleasure is

Before we discuss what hedonism is, we should first clarify the concepts of pleasure and pain. In the following I will mainly concentrate on pleasure, but we may assume a similar explanation applies, *mutatis mutandis*, to the concept of pain.

Sidgwick's explanation of the concept of pleasure appears in several places in *ME*. The first thing we should note is that pleasure is a type of feeling.

> First, I will concede that pleasure is a kind of feeling which stimulates the will to actions tending to sustain or produce it, – to sustain it, if actually present, and to produce it, if it be only represented in

> idea –; and similarly pain is a kind of feeling which stimulates to actions tending to remove or avert it.
>
> (*ME* Bk. 1 Ch. 4 pp. 42–3. See also *ME* Bk. 2 Ch. 2)

This stimulus to will is called desire or aversion.[2] Desire is explained as a felt stimulus or impulse to an action that tends to realize what is desired (*ME* p. 43 fn. 2), and we can rephrase it as a conscious impulse that aims at a certain object or at an action to obtain it.

Thus pleasure is a kind of feeling that arouses a desire to maintain or realize it. This feeling becomes an object to be desired, and when this feeling has been generated we can say that the desire to realize it is satisfied at that moment (though the desire to maintain that feeling keeps on working). Thus pleasure is essentially related to desire. However, pleasure itself is a feeling (see, for example, *ME* pp. 43–4) and not a synonym for desire. In our ordinary life we also desire many other things than pleasure, including various things or states of affairs, ideals, knowledge, etc. Since pleasure is related only to a special desire that aims at the pleasant feeling itself, we cannot equate pleasure with desire. Nor can we equate pleasure with *the object* or *the satisfaction of desire* (i.e., the occurrence of a state of affairs which fulfills one's desire). This is partly because obtaining the desired object or fulfilling the desire is not always accompanied with a *feeling* of satisfaction,[3] and partly because the actual objects of our desire are not limited to feelings. Pleasure does not account for *all* the objects of our desire; it is a subset of various desired or desirable objects, which are grouped under the common head of 'feelings'. Pleasure is a kind of *feeling* that would be desired by the person who is experiencing it, and we cannot simply define it as 'what is desired' or as 'the situation that fulfils a person's desire'.

To be noted further, pleasure cannot be defined even as a feeling that *actually arouses* our desire. Pleasure usually arouses desire, but it may not arouse one's desire when one envisions it only as an abstract idea; in addition, we may often suppress our desire for a certain pleasure when we consider it impossible to obtain. Pleasure is a kind of feeling that *may* become the object of our desire, but we cannot simply call it an *actually desired* feeling. Rather, it is a '*desirable*' feeling, as explained later.

The second point we should remember is that Sidgwick understands the concept of pleasure in a very wide sense. This extension of the concept of pleasure is clearly suggested in the following passages from *ME*.

> To be clear, then, we must particularise as the object of Self-love, and End of the method which I have distinguished as Egoistic Hedonism,

> Pleasure, taken 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.
>
> (*ME* Bk. 1 Ch. 7 Sec. 2 p. 93. We should understand his term 'satisfaction' not as an objective situation in which what one desired has actually happened, but as a satisfied *feeling* in a person's mind, for Sidgwick clearly explains that pleasure is a kind of feeling.)

> [W]hen I reflect on the notion of pleasure, – using the term in the comprehensive sense which I have adopted, to include the most refined and subtle intellectual and emotional gratifications, no less than the coarser and more definite sensual enjoyments.
>
> (*ME* Bk. 2 Ch. 2 Sec. 2 p. 127)

What Sidgwick imagines with the term pleasure is a very comprehensive notion, which includes not only sensual or physical pleasure but also mental pleasure and enjoyment, and which also encompasses calm and peaceful gratifications as well as acute and excited sensations. *Whatever* feeling that would become the object of desire Sidgwick calls pleasure.

One point becomes clear here. For Sidgwick, happiness is equivalent to pleasure, or it consists of pleasures. Some may insist, however, that happiness is apparently different from pleasure or from a mere aggregation of pleasures (see, for example, Shionoya 1984, pp. 392–3). Such people may claim that one should not aim for pleasure, which is sensual and temporary, but for happiness, which is a more profound satisfaction that can be obtained when one contemplates one's life in the long run. However, Sidgwick would include such 'happiness' into the wide category of pleasure, for it is certainly a kind of satisfied feeling that can be obtained when one contemplates one's life.

The two passages cited above appear in Sidgwick's explanation of egoism (egoistic hedonism), but we may regard these as a description of pleasure in general. In Book I, Chapter 3 of *ME*, where Sidgwick deals with pleasure in general, he similarly refers to a wide variety of pleasure, from the pleasure of eating, through that of contemplative, investigational and creative activities, to extremely refined pleasures.

So far we have roughly explained pleasure as 'a kind of feeling that may be desired'. We have not, however, clarified a very important point about the concept of pleasure. The third point Sidgwick suggests is that

pleasure can only be directly known to the individual who experiences it. More precisely, pleasure is only directly known to an individual during the moment of feeling (*ME* pp. 122, 128, etc.). Perhaps we may regard this as commonly understood. All feelings are subjective, and other people cannot directly feel them. Even in one and the same person's mind, a feeling changes and vanishes over time, and once it fades away he cannot directly feel it again. He is also unable to experience his future feelings. People cannot directly feel another person's pleasure, nor they can precisely determine general rules about how and when pleasure might occur. How pleasure is generated differs from person to person, and from time to time. Some may greatly enjoy a certain television program, but others may feel disgusted at its silliness. Some may feel delighted when they eat a lavish dinner, whereas others may feel sickened by it. A pleasant feeling is not determined merely by certain objective conditions; after all, it is a personal and transient feeling. The only person who can precisely know whether a certain pleasant feeling is present, and how it feels, is the individual who is feeling it at the time. In all the cases just described, however, we would commonly admit that the kind of feeling that involves desire (such as enjoyment, joy, delight, etc.) can be called pleasure, and that the kind of feeling that involves aversion (disgust, illness, etc.) can be called displeasure or pain. It is an individual at one point in time who feels pleasure, but the meaning of the concept of pleasure is common to us all. This is why we can use the common term pleasure.

7.2.2 Definition of pleasure for quantitative comparison

A specific definition of pleasure appears in Book II, Chapters 2 to 3 of *ME*, where Sidgwick presents arguments about what he calls 'empirical hedonism'. Empirical hedonism is one of the hedonistic methods that prescribe how to promote pleasure, provided that pleasure is good. Other types of hedonism, 'objective' and 'deductive' hedonism (see 2.2 of the present book), claim that it is difficult to precisely foresee and measure each particular pleasure and that therefore we should use indirect methods to choose actions that are likely to bring about pleasure. On the other hand, empirical hedonism claims that we should reflect on our own experiences to foresee, measure and maximize pleasure that will result from an action. In egoistic hedonism, this pleasure is confined to one's own pleasure, and in universalistic hedonism, it encompasses all people's pleasure. Thus, in empirical hedonism, we need to find a clear definition of pleasure, and especially the definition that can be used to make quantitative comparisons of pleasure.

In Book II, Chapters 2 and 3 of *ME*, where the definition of pleasure appears, Sidgwick primarily considers a comparison of one's own pleasure. This is because the main topic of *ME* Book II is egoistic hedonism. Therefore, for the time being, I will consider the cases in which one makes comparisons about one's own pleasures, leaving until later the topic of interpersonal comparisons of pleasures. We should note, however, that the definition of pleasure that is presented in Book II, Chapters 2 and 3 is valid also as the definition of pleasure in general.

Now, we regard various states of mind as pleasure, but what they all have in common is that they are feelings that normally excite desire when actually felt or imagined. Thus the first proposal that comes to our mind is to define pleasure as the feeling that one actually desires when one feels or precisely imagines it (in the sense that one represents to oneself), and to measure the size of such pleasure by the intensity of desire in the person who feels it or represents it to oneself.

However, as far as ordinary judgments are concerned, it seems that the intensity of our actual desire for pleasure is not precisely in proportion to the size of the pleasure (*ME* p. 125). First, when we take pleasure in rest or basking in the sun, a person's desire for it is not always manifest but becomes explicit when that pleasure is terminated. We can say that in such cases one has the desire only potentially and implicitly; but one would probably claim, when asked, that one desires or prefers it. Yet even in cases where a person clearly feels desire, the actual strength of his desire does not necessarily correspond with the size of his pleasure. This is partly because a person sometimes suppresses his own desire when he knows that he cannot attain that pleasure, and partly because he may not strongly desire a future pleasure, even though he knows that it could bring great pleasure. Thus we cannot call pleasure the feelings that are *actually desired* by the person who feels it. Sidgwick proposes to explain pleasure as 'desirable' feelings:

> [T]he only common quality that I can find in the feelings so designated seems to be that relation to desire and volition expressed by the general term 'desirable', in the sense previously explained.
>
> (*ME* Bk. 2 Ch. 2 Sec. 2 p. 127)

The term 'desirable' here seems to have the same connotation as when it was used in the definition of 'good'. Then, 'desirable' means not 'what is actually desired' nor 'what ought to be desired' but 'what would be desired, with strength proportioned to the degree of desirability, if it were judged attainable by voluntary action, supposing the desirer to

possess a perfect forecast, emotional as well as intellectual, of the state of attainment or fruition' (*ME* pp. 110–1. See also 4.3.4 of the present book). Thus pleasure, or a desirable feeling, is the feeling that *would* be desired (with the strength proportional to the degree of desirability) if the following two ideal conditions are met: (1) that one judges that one can maintain or produce that feeling by one's will, and (2) that one fully foresees, both emotionally and intellectually, the mental state in which this feeling is present.

This being so, the size of one's pleasure seems to be in proportion to the intensity of one's *ideal* desire for that feeling, which corresponds to the degree of the desirability of this feeling. If a certain feeling would be desired more strongly than another when the above two conditions are met, it is the *greater* pleasure; and the feeling that I would desire most under such ideal conditions is called the *greatest* pleasure. Here, for feeling A to be desired more strongly than feeling B is for A to be preferred to B, meaning that the will to choose the action which will realize A is more strongly stimulated. To rephrase the above argument, then, the feeling that would be preferred to all other feelings under the above two conditions is the greatest pleasure. This comparison of pleasure can be done by the person who foresees all attainable pleasures and vividly imagines them in his mind.

One may suggest that, since pleasure is defined not only by the term *desirable* but also constitutes a *feeling*, the comparison of pleasures can be done in terms of the vividness of the feeling or the intensity of its sensation. However, Sidgwick emphasizes that 'we must be careful not to confound intensity of pleasure with intensity of sensation' (*ME* Bk. 1 Ch. 7 p. 94. See also 3.1.1 of this book). His claim that the comparison of pleasures must be done not by the intensity of sensation but by the intensity of ideal preferences can be regarded as valid for the following two reasons. First, we are now dealing with pleasure in the context of considering how to make an ethical choice of action. Generally, a person's choice depends not on his sensation but on his preference, and his best ethical choice would be based on his well-considered preference. When a person makes certain choices (to donate to some nonprofit organization, to steal something from others, or whatever else he chooses), he does so not because he feels a certain acute *sensation* but because his overall *preference* urges him to do so. Therefore 'desirability' or 'preferability' is essential to our consideration of ethical decision-making, whereas sensation is not. Second, the distinctive feature which all pleasant feelings have in common, and which distinguishes pleasures from nonpleasant feelings, is their relationship to desire or

will, which is called 'desirable'. The existence of sensation is not what distinguishes pleasure from other feelings because every kind of feeling is accompanied by a certain sensation; in contrast, whenever a feeling is (or *would* be) accompanied by one's desire for it, we can call it pleasure. Furthermore, we do not think that a feeling with stronger or weaker sensation is always greater or lesser pleasure. It is when one's stimulus to realize and maintain such a feeling is strong, under the ideal conditions described above, that we call it a great pleasure, and when one's stimulus to avert and eliminate it is strong we call it a great pain. If we are to compare the size of pleasure and pain, the only method to do so is to compare the intensity of ideal preferences for those feelings.

Nevertheless, when we imagine pleasure only as an abstract idea, we cannot precisely measure its desirability, that is, the intensity of desire that we *would* have if we could perfectly forecast the state of its attainment or fruition. This is because we are actually unable to make such a perfect forecast. Our imagination for future pleasure is often deficient, and our past pleasure loses its luster over time. Nevertheless, we commonly admit that pleasure is directly known only to the individual at the time of experiencing it. In the intellectual being who is experiencing it, pleasure is fully recognized and felt in mind. Here Sidgwick's second ideal condition is certainly satisfied in that one fully imagines, both emotionally and intellectually, the mental state in which this feeling is present. Therefore, the strict measurement of pleasure is being done by the very individual at the moment of experiencing it, because only he can judge the desirability of the pleasure he is experiencing. However, as already explained, even in the very person who is experiencing a pleasant feeling, his desire for that feeling may not be distinctly aroused. When a person is feeling a pleasure, his desire to *realize* that pleasure is already satisfied at that moment, and his desire to *maintain* it may be unconscious (as in the case of pleasure in rest or sleep). In addition, when he judges that he cannot maintain this feeling by his own effort, he may suppress his desire for it. Still, we can say that the individual who is experiencing a pleasure does at least implicitly apprehend this feeling as 'desirable'. Thus Sidgwick proposes the following definition of pleasure.

> I propose therefore to define Pleasure – when we are considering its 'strict value' for purposes of quantitative comparison – as a feeling which, when experienced by intelligent beings, is at least implicitly apprehended as desirable or – in cases of comparison – preferable.
>
> (*ME* Bk. 2 Ch. 2 Sec. 2 p. 127)

This definition, however, is still insufficient. A person who is experiencing a pleasure can also misjudge the desirability of that feeling, if his judgment is mixed with other considerations such as the conditions that bring about the feeling in question, the circumstances that concur with the feeling, and the subsequent influences of such feeling on himself or on other people. For example, one may judge that the pleasure of taking an illegal drug is 'undesirable' because such a judgment is easily mixed with our concern that it is illegal or that taking it will have undesirable physical or psychological aftereffects; but we may admit that the drug certainly brings about a 'great pleasure' if its pleasantness is solely considered, completely separately from its circumstances, consequences and all other factors. Sidgwick believes that this latter comparison is what we need for the precise evaluation of the desirability of pleasant feelings. If we are to purely measure the size of pleasure and pleasure alone, we have to consider only that feeling, removing all other factors than that feeling. Then, if an individual measures the desirability of his own feeling at the time of feeling it, separately from any other factors, no one would be able to deny his evaluation. Thus Sidgwick reaches the final version of the definition of pleasure.

> Let, then, pleasure be defined as feeling which the sentient individual at the time of feeling it implicitly or explicitly apprehends to be desirable; – desirable, that is, when considered merely as feeling, and not in respect of its objective conditions or consequences, or of any facts that come directly within the cognizance and judgment of others besides the sentient individual.
>
> (*ME* Bk. 2 Ch. 3 Sec. 1 p. 131)

Sidgwick's arguments so far can be summarized as follows. Pleasure is a feeling that a person would, at least implicitly, desire to maintain or realize if (1) he fully imagines (i.e. represents to himself) the state in which this feeling is present, (2) considers this feeling apart from all other factors, and (3) judges that it is possible for him to attain or realize that feeling. The only person who can fully imagine the state in which a pleasure is present is the individual who is experiencing it. Therefore, the accurate measurement of 'desirability' of a certain pleasure is likely to be done by asking the person who is experiencing it how strongly he would desire that feeling when he solely considers that feeling apart from other factors, and when he believes that it is possible to maintain or realize it. This is the way to determine the 'strict value' of pleasure proposed by Sidgwick. If we can determine such value, then we can compare different pleasures by their size (see *ME* p. 127).

However, there is a problem with this comparison of pleasures if the person at the moment of experiencing pleasure is the *only* one who can precisely measure its desirability or preferability. This is the problem of whether we can precisely compare the desirability of pleasures that occur at different points in time. The argument just described suggests that such a comparison cannot be precisely done, for it assumes that a person at one point in time cannot accurately measure pleasure at another point in time.[4]

Sidgwick takes this difficulty seriously, but still claims that in ordinary life we undoubtedly compare the strength of pleasures and pains at different times, and that we will continue to do so no matter what problem may arise regarding that comparison. According to Sidgwick, the best attainable method for a comparison of two future pleasures would be for an individual to picture in his present mind as precisely as possible, making full use of his experience and observation, how he will feel each pleasure at the time of enjoying it (see *ME* pp. 140, 150, 195). In the case of comparing past and future pleasures, he has to recall, on the one hand, how he felt the past pleasure at that moment, and on the other hand imagine how he will feel the future pleasure before he compares their desirability. When we believe that we can make quantitative comparisons of pleasures, we have to assume that (1) feelings can be compared with each other, clearly enough for practical purposes, by considering the desirability of the feeling to be measured by the person who is experiencing it. Furthermore, in order for *me* to compare my pleasures at different times, I have to assume that (2) the degree of their desirability can be known to myself at the time of experiencing them as having a certain definite value. Such a measurement can be expressed as a kind of numerical value that could be understood by myself at other points in time (*ME* pp. 129, 131). We could deny these assumptions, of course, but probably most of us have these presuppositions when comparing our own pleasures.

Sidgwick provides the above explanation about pleasure in Book II of *ME*, where he presents arguments on egoistic hedonism. The comparison of pleasure just described is intended to be a comparison of one's own pleasures. However, we assume that the definition of pleasure should be valid throughout *The Methods of Ethics*. In Book III, Chapter 4 of *ME*, Sidgwick states that others' happiness should also be understood as what each individual would judge to be desirable (*ME* p. 240). Here he obviously considers happiness, or pleasure, in the same sense explained in his arguments on egoistic hedonism. Additionally, in his description of utilitarianism in Book IV of *ME*, Sidgwick states that the

notion of the greatest happiness clarified in Book II can be applied not only to egoistic hedonism but also to universalistic hedonism. He also claims that, though it is even more difficult to compare different people's happiness, empirical hedonism is the clearest method to attain maximum general happiness (*ME* Bk. 4 Ch. 1 p. 413 f.; Bk. 4 Ch. 4 p. 460). In short, whether he considers one's own pleasures or peoples' pleasures, he intends to measure and compare them by the same criterion of 'desirability that would be judged by the individual at the time of experiencing it'. Therefore, universalistic hedonism has to have similar assumptions as egoistic hedonism, that is (1) that pleasure and pain have different degrees of intensity, which can be shown as having definite values in proportion to their desirability, and (2) that each of such values can be known not only to the individual who is experiencing them but also can be understood to some extent by other individuals who might compare these various feelings. Thus one can attempt to make quantitative comparisons of one's own and others' pleasures using this common criterion (*ME* p. 413). Since there are few logical grounds to support these assumptions, we could deny them; and there is no guarantee that we can precisely compare one's own or other people's pleasures. I will discuss this problem of hedonistic comparison in Chapter 10 of this book. In the present context, where we are dealing with hedonism, it is sufficient that we accept two points: first, that the final judgment of a pleasure's desirability rests with the individual who is experiencing it, and second, that we commonly admit that the degree of such pleasure or pain is discerned clearly enough for practical purposes. According to Sidgwick, we have to evaluate pleasure, whether it is one's past/future pleasure or someone else's pleasure, by its desirability judged by the person who is experiencing it. The assumption that such desirability can have some definite value is deniable, but it is what we must presuppose if we are to compare pleasures.

To sum up the points of the concept of pleasure: one, Sidgwick uses the term pleasure in a very wide sense; two, pleasure is a type of feeling that is expressed as a *desirable* feeling; three, it is commonly admitted that the size of a pleasure, if measurable, is to be determined by the degree of that feeling's desirability from the viewpoint of the individual who is feeling it. Assuming that such desirability takes a certain definite value, we can make a comparison of pleasures, and we usually make such a comparison (even though it may be a rough one) on this assumption.

Now that we have clarified Sidgwick's concept of pleasure, we now have to ask how this concept of pleasure is related to the concept of

good. This is the essential question of hedonism. Before we look into this question, however, we need to distinguish the different types of hedonism.

7.2.3 Pleasure and good: Psychological *vs* ethical hedonism

'Hedonism' can mean any one of several different claims. The first distinction we should make is between psychological and ethical hedonism. *Psychological hedonism*, or *hedonism on psychological fact*, claims to describe the psychological fact that the sole object of our actual desire is pleasure, or that our will is always determined by actual or foreseen pleasures and pains. *Ethical hedonism*, or *hedonism on good to be pursued*, claims to propose a value judgment that pleasure is the sole ultimate good and the ultimate end of a right action (see *ME* pp. 40, 388). We may further distinguish sub-categories for each kind of hedonism, as cases in which an individual's own happiness (egoistic pleasure) is solely considered and those in which people's general happiness (universal pleasure) is considered. Thus we can obtain the following table of classification.[5]

(I) ***Psychological hedonism:*** The object of each individual's actual desire is always pleasure.
  (i) ***Psychological egoistic hedonism:*** The object of each individual's actual desire is always his own pleasure.
  (ii) ***Psychological universalistic hedonism:*** The object of each individual's actual desire is always people's pleasure.

(II) ***Ethical hedonism*:** Pleasure is the sole ultimate good, and the ultimate end of a right action.
  (iii) ***Ethical egoistic hedonism:*** One's own pleasure is the sole ultimate good, and the ultimate end of a right action.
  (iv) ***Ethical universalistic hedonism*:** People's pleasure is the sole ultimate good, and the ultimate end of a right action.

We should note that the term pleasure here is used in a very broad sense, as Sidgwick has defined it. For example, when a person feels satisfied or delighted by doing beneficence to others, his satisfied or delighted feeling is classified as his own pleasure. Provided that pleasure is to be understood in a wider sense, psychological egoistic hedonism (i) claims that each individual *does* pursue one's own pleasure or some kind of satisfied feeling, and ethical egoistic hedonism (iii) claims that each individual *ought to* pursue one's own pleasure or some kind of satisfied feeling. The claim that one ought to pursue not only one's

own pleasure but also others' satisfied or delighted feelings, or people's general pleasure, is ethical universalistic hedonism (iv).

Of these four subcategories, Sidgwick tends to label (i) as 'psychological hedonism', and (iv) as 'ethical hedonism'.[6] For the sake of argument, however, I will use my own classification, as stated in the table above. Needless to say, *ethical* egoistic hedonism (iii), or egoistic hedonism *on good to be pursued*, is the basic principle of egoism (i.e. egoistic hedonism in Sidgwick's term), and *ethical* universalistic hedonism (iv), or universalistic hedonism *on good to be pursued*, is the basic principle of utilitarianism (i.e. universalistic hedonism). I might add that psychological universalistic hedonism (ii) is usually regarded as untrue. People's general happiness or universal pleasure means the overall happiness of *all* the members of a certain group. Thus psychological universalistic hedonism is not the claim that everyone always desires someone else's pleasure but that everyone always desires the overall pleasure of all parties. No one would admit that this is true.

Some utilitarian thinkers, such as Jeremy Bentham, have supported psychological egoistic hedonism and ethical universalistic hedonism at the same time, claiming that pleasure is the only thing that one actually pursues *and* the only ultimate good that one ought to pursue. Sidgwick finds these two factors, psychological egoistic hedonism and ethical universalistic hedonism, in J. S. Mill as well (Preface to *ME*6). Such utilitarians sometimes claim that ethical universalistic hedonism can be 'proved' by an argument based on psychological egoistic hedonism. This line of argument is often attributed to Mill's explanation in his *Utilitarianism*.[7] However, Sidgwick takes a different tack on hedonism. What is remarkable about Sidgwick is that he endeavors to support ethical hedonism without upholding psychological egoistic hedonism. Moreover, he endorses ethical hedonism bypassing a choice between ethical egoistic hedonism and ethical universalistic hedonism; his strategy is to defend ethical hedonism as something that can be applied equally to ethical egoistic and universalistic hedonism. Utilitarianism as universalistic hedonism is not supported unless ethical hedonism is combined with the Principle of Rational Benevolence.

Sidgwick points out three errors that we frequently make concerning the correlation between any two of the four subcategories classified above. One is the confusion of *psychological* egoistic with *ethical* egoistic hedonism. The second is an erroneous attempt to prove ethical universalistic hedonism by means of psychological egoistic hedonism. The third is the fallacy of psychological egoistic hedonism and psychological hedonism in general. Sidgwick claims that the object of our actual desire is *not always* pleasure.

As for the first error of confusion between psychological and ethical egoistic hedonism, Sidgwick flatly denies that there is any logical or inevitable relationship between the former, factual statement, and the latter, ethical judgment (*ME* pp. 40–1). This reflects a famous argument that factual and evaluative judgments fundamentally differ from each other. Sidgwick admits, however, that, if it is certain that the end of a person's action is always his own pleasure (or absence of pain), we cannot tell him to pursue another end contrary to this psychological law. We may similarly admit that, if the proposition that 'a person always pursues people's universal pleasure' were true, we would not be able to accept a judgment other than that 'people's pleasure is the ultimate good'.

However, Sidgwick contends that it is impossible to prove ethical universalistic hedonism based on psychological egoistic hedonism. Even if we admit that ethical *egoistic* hedonism may be derived from psychological egoistic hedonism, this does not show the truth of ethical *universalistic* hedonism; after all, it merely endorses egoism. Moreover, even if we admit that ethical universalistic hedonism may be derived from psychological universalistic hedonism, most of us will regard psychological universalistic hedonism as false, and psychological egoistic hedonism will not produce psychological universalistic hedonism. According to psychological egoistic hedonism, each individual *always* seeks his own pleasure; and it does not follow from this proposition that each individual *always* seeks people's overall pleasure. When each person seeks his own pleasure, people's respective desires are directed to different parts of people's overall pleasure, and a hodgepodge of such desires cannot generate a collective desire for people's overall pleasure. Thus, we cannot succeed in proving ethical universalistic hedonism based on psychological egoistic hedonism via psychological universalistic hedonism (*ME* p. 388).

Some writers claim that we can find another way to support ethical universalistic hedonism based on psychological egoistic hedonism (this point is suggested by Uchii 1988, p. 203 ff.). Their strategy is to show that an individual inevitably comes to accept ethical universalistic hedonism if he realizes he needs to respect other people's pleasure in order to pursue his own pleasure. This line of argument derives ethical egoistic hedonism from psychological egoistic hedonism, *without* asserting psychological universalistic hedonism, and then claims that a refined ethical egoistic hedonist would adopt ethical universalistic hedonism. To be precise, this argument proceeds as follows. If everyone actually seeks his own pleasure, all individuals will not want their own

pleasure to be ignored or prevented by other people. Thus, in order for us to fulfill our own desire, we must accept the evaluative judgment that anyone's own happiness or pleasure is good for oneself. So we must admit that we ought to respect other people's happiness or pleasure as well as our own. When all of us have the same desire to pursue our own happiness without being hampered by others, our best policy would be to coexist with others' similar desires by avowing that people's general happiness is the sole desirable good.

However, Sidgwick goes further to show that the major premise in this line of reasoning collapses, by denying the very claim of psychological egoistic hedonism (*ME* Preface to *ME*6; Bk. 1 Ch. 4 Sec. 2). We do not always act out of the conscious pursuit of our own pleasure (or avoidance of pain). For example, it is not certain that one's appetite really targets the satisfaction one gets from eating; and there are many other desires which do not consciously target the agent's own pleasure. Such desires include disinterested altruistic ones, for example, the desire to dive into the sea to save a drawing child, and ones that target no one's pleasure, for example, the desire to realize a certain ideal or to pursue truth, or the destructive desire to overspend, to use drugs or to hurt someone. We often have such desires even though we know that to act according to those desires will not bring about pleasure to ourselves, or that it may result in self-sacrifice or self-destruction. We experience conflict between the desire for something other than our own pleasure on the one hand, and the desire for our own pleasure on the other hand; this is what causes us to worry about taking action.

The fact that our desire does not always target our own pleasure also becomes clear when we consider one's desire for achievement. According to Sidgwick, we can distinguish the desire to achieve a certain goal from the desire for pleasure that can be obtained from such achievement. Suppose that we are going to play a game in which victory and defeat will be clearly determined. Players may not initially have the desire to achieve victory. They may simply be enjoying the process of the game. In order to fully enjoy playing the game, however, it would be favorable to have a desire to win the game. The more both players aim for victory, the more exciting the game becomes; and, as a result, both players can obtain maximum enjoyment from the game. Thus a player's desire for victory can be newly generated *because* he pursues the pleasure of playing the game. The direct target of this newly created desire is, however, not the pleasure of playing the game but victory. Its conscious target is not even the *pleasure* of attaining victory. True, the stronger this desire for victory becomes, the more pleasant its achievement may

seem. Sidgwick suggests, however, that we cannot call such a *pleasure* of achieving victory the 'target' of this desire. The target of this desire is achieving victory. If this is correct, this is another kind of desire that does not consciously aim for pleasure.

Sidgwick adds one more point to his argument against psychological hedonism. He claims that, in undertaking a certain activity, we often get the greatest attainable pleasure when we do not consciously pursue it, rather temporarily allowing other impulses to perform the activity in question. For example, we know there are pleasures felt during creative activities or intellectual inquiries. Such pleasures can be fully obtained when we do not directly pursue them but rather forget ourselves while we are completely absorbed in creating wonderful works or inquiring into profound truths. Ironically, we cannot fully enjoy such pleasures when our conscious desire to gain pleasure become so dominant that they impede us from performing activities to our full potential. Sidgwick calls this the fundamental paradox of hedonism (*ME* pp. 49, 136, 403).[8]

Thus Sidgwick concludes that, as far as our actual actions and feelings are concerned, our conscious impulses are not always directed to our own pleasure (or avoidance of pain) (*ME* p. 52).

Two objections to this conclusion can be made, however, by the supporters of psychological-egoistic hedonism. The first objection is that, though we do not always consciously pursue our own pleasure, we do so *unconsciously*. The second is that, though all impulses originally target our own pleasure, by association we have come to have impulses toward other things. Even though we admit these objections, Sidgwick argues, we cannot use them to support ethical hedonism. As for the first objection, 'for a person to unconsciously pursue his own pleasure' only means that the person himself is unaware of it but others may judge him to be pursuing his own pleasure. This point is irrelevant for the claim of ethical hedonism. Nor is the second objection sufficient as the basis of ethical hedonism. At any rate we must admit that *now* we are not consciously pursuing our own pleasure alone. Whatever the origin of desire, we sometimes do wish for other people's happiness, and we sometimes immerse ourselves in our work without considering whether it will bring about our own pleasure. As long as we admit this, these two objections are not convincing arguments for adopting pleasure as the end to be pursued (*ME* p. 52).

Sidgwick thus admits that we sometimes desire other things than pleasure. He also concedes that the fulfillment of a desire does not always result in pleasure. If we admit these arguments, we cannot prove

ethical universalistic hedonism by appealing to psychological egoistic hedonism. According to Sidgwick, this does not mean that we cannot make a case for ethical hedonism. However, Sidgwick never attempts to prove psychological universalistic hedonism, which is usually regarded as untrue; he does not support psychological hedonism – whether egoistic or universalistic – in the first place. We know that we often have desires for an ideal or for beauty, which may not be identified with pleasure. The claim that we always consciously or unconsciously pursue someone's pleasure has no power to prove ethical hedonism. For Sidgwick, it is sufficient if we can directly establish ethical hedonism by reflecting on our philosophical intuition, regardless of the psychological facts about our ordinary actions.

Then, what do we mean by the claim 'the ultimate good is pleasure and pleasure alone'?

Let us reconsider the relationship between pleasure and good. Good is what is desirable, and pleasure is a desirable feeling. Therefore, the proposition that pleasure is a good is analytically true, but the proposition that pleasure is the *sole* good is not. In reality, what is usually called good is not limited to pleasure. Admitting that we judge various other things as good, and admitting that the proposition that pleasure is the sole good is a synthetic one, Sidgwick argues that the feeling of pleasure is the sole ultimate good.

The ultimate good is what is desirable in itself, as explained in 4.3.1 of this book. Pleasure is a feeling that is apprehended as desirable. Sidgwick explains how to 'prove' ethical hedonism, that is, the claim that pleasure is the sole ultimate good, as follows:

> It should be observed that if this definition of pleasure be accepted, and if, as before proposed, 'Ultimate Good' be taken as equivalent to 'what is ultimately desirable', the fundamental proposition of ethical Hedonism has chiefly a negative significance; for the statement that 'Pleasure is the Ultimate Good' will only mean that nothing is ultimately desirable except desirable feeling, apprehended as desirable by the sentient individual at the time of feeling it.
>
> (*ME* Bk. 2 Ch. 2 Sec. 2 p. 129)

However, the objects of our actual desire, and the objects that we usually judge to be desirable, are not limited to pleasure. Sidgwick thus endeavors to support ethical hedonism by arguing that, though we usually judge various things to be desirable, *if we carefully reflect on our own intuition* we will agree that pleasure is the only thing that is truly

desirable in itself and not as a means to other things. This argument is called the 'proof' of ethical hedonism.

7.2.4 Proof of ethical hedonism

The proof of ethical hedonism is an argument which attempts to show that the ultimate good is pleasure, and pleasure alone. (To simplify, henceforth I will often call ethical hedonism simply 'hedonism'.)

According to Sidgwick, ethical hedonism is a truth that can be apprehended by intuition.

> If Hedonism claims to give authoritative guidance, this can only be in virtue of the principle that pleasure is the only reasonable ultimate end of human action: and this principle cannot be known by induction from experience. Experience can at most tell us that all men always do seek pleasure as their ultimate end (that it does not support this conclusion I have already tried to show): it cannot tell us that any one ought so to seek it. If this latter proposition is legitimately affirmed in respect either of private or of general happiness, it must either be immediately known to be true, – and therefore, we may say, a moral intuition – or be inferred ultimately from premises which include at least one such moral intuition.
>
> (*ME* Bk. 1 Ch. 8 p. 98)

Then, why do we need to 'prove' such an intuitive truth? This is partly because we do not always understand an intuitive truth with sufficient clarity, and partly because the intuitively apprehended truth is only a *prima facie* truth (see 3.2 of this book). Moreover, good and pleasure are certainly not synonymous, and what we usually judge to be good is not limited to pleasure. In short, that the ultimate good is pleasure is a synthetic proposition. Therefore, we need some reasoning by which we become convinced that pleasure is the only ultimate good.

Interestingly, Sidgwick's proof of hedonism takes an inductive method of ascertaining this intuitive truth. This is not a self-contradictory strategy, as suggested in my explanation of the relationship between induction and intuition in 3.2.2 of the present book. As the above citation suggests, the content of an intuitive truth (that ethical hedonism is true) cannot be known by induction from experience alone. However, our argument would be consistent if, *after* we collected several candidates for the ultimate good by induction from experience, we ask ourselves which of these candidates we regard as a truly valid ultimate good *by appealing to our own intuition*. Then, Sidgwick's proof of hedonism is

the argument that, though we usually desire various things including pleasure and nonpleasure, and though we often judge many nonpleasant things to be good, if we use our reflective intuition to consider the question of what is truly desirable in itself, our answer will be nothing but pleasant feelings.

In a sense, this proof of ethical hedonism appeals to a fact about what we actually do, namely, the fact that we actually judge pleasure to be the sole ultimate good when we use our reflective intuition. However, this line of argument is utterly different from the proof of ethical hedonism based on psychological hedonism. Sigwick's proof of ethical hedonism appeals not to the (false) fact that we always consciously or unconsciously seek pleasure but to the anticipated fact that *on reflection* we will judge pleasure to be the ultimate good. His proof rests not on a psychological fact that governs our everyday actions but on our value judgment that we make when we reflect on what is considered to be good in itself.

We should also note that what Sidgwick is about to prove is ethical hedonism *in general*, which is not limited either to ethical *egoistic* hedonism or to ethical *universalistic* hedonism. By proving ethical hedonism Sidgwick obtains everything he needs to construct the foundation of utilitarianism; but this proof itself is the argument to show that what our reflective intuition judges to be the ultimate good is always *someone's* pleasure. Sidgwick does not claim that the ultimate good is one's own pleasure alone, nor it is people's universal pleasure or general happiness. Ethical hedonism in this general sense only claims that, whether 'an individual's good on the whole' or 'people's general good' is concerned, what constitutes good is nothing but someone's pleasure. Sidgwick's proof of hedonism is unique in this sense, too. He argues not that we always seek our own pleasure, nor that our intuition will judge our own pleasure to be the sole ultimate good, but that our intuition will judge *pleasure in general* to be the sole ultimate good.

If this proof of hedonism is successful, and if we admit that ethical hedonism is a universal truth, the ultimate criterion for comparing different 'goods' would be pleasure. When we discussed the differences between pleasure and good in 4.3.3 of this book, we saw that, according to our ordinary judgments, the goodness of a certain good thing corresponds only to *a certain type* of pleasure that it brings about, and the degree of its goodness is not proportional to the amount of pleasure which is generated. This disagreement about good and pleasure in our ordinary judgments, however, will be revised on reflection, after

we are convinced that pleasure is the sole ultimate good. That is, we must admit that an appropriate evaluation of the goodness of a certain object should be based on the consideration of the amount of pleasure produced by that object.

Some may wonder how such a personal feeling as pleasure can become a 'universally valid' criterion for the comparison of good. As stated in the analysis of the concept of pleasure, pleasure is directly known only to the individual who experiences it. However, we have a common conception of pleasure, and most of us would admit that the criterion for comparing different pleasures is the desirability of those pleasures, which are judged by the individual at the time of feeling them. The claim that pleasure is the universal criterion for comparing different 'goods' only means that (i) we should universally admit that pleasure is to be the sole factor in evaluating good, and that (ii) we should all admit that the common criterion for comparing pleasures should be applied when comparing various 'goods'. There is no detailed explanation about the relationship between these two criteria to evaluate pleasure and good in *The Methods of Ethics*, but the following passage, especially (2), should be noted.

> [I]n affirming Ultimate Good to be Happiness or Pleasure, we imply (1) that nothing is desirable except desirable feelings, and (2) that the desirability of each feeling is only directly cognisable by the sentient individual at the time of feeling it, and that therefore this particular judgment of the sentient individual must be taken as final on the question how far each element of feeling has the quality of Ultimate Good. Now no one, I conceive, would estimate in any other way the desirability of feeling considered merely as feeling.
>
> (*ME* Bk. 3 Ch. 14 Sec. 4 p. 398)

Now let us look into Sidgwick's proof of ethical hedonism. His entire arguments take the following structure (*ME* Bk. 3 Ch. 14)[9]:

Step one: Argument of elimination. If we carefully examine all conceivable candidates for the ultimate good and narrow them down by eliminating unsuitable ones, the only candidate that is left as truly desirable is 'desirable feeling', that is, pleasure.

Step two: Appeal to an individual's intuition and commonsense. If we seriously reflect on our intuition and commonsense, we will be convinced that they support hedonism. We can also use this reflection to answer to the common objections against hedonism.

Step three: Argument from practical need. We cannot find any other coherent systematic theory that meets our practical needs.

Step one is virtually the central argument to 'prove' ethical hedonism. In it Sidgwick develops a philosophical analysis based on empirical observations, and draws the conclusion that, when he appeals to his own reflective intuitions, he cannot regard anything but pleasure as the ultimate good. Step two supports the validity of the conclusion of Step one by appealing to other people's personal intuitions and commonsense. Step three is a supplemental support for ethical hedonism, referring to the realistic demand that we have no choice but to accept ethical hedonism if we are to coherently make practical decisions. Throughout the three arguments, Sidgwick consistently refers to what our refined intuition would judge to be the ultimate good. Ethical hedonism as an intuitive truth is established not by being derived from the empirical fact of psychological hedonism but by means of examining whether ethical hedonism satisfies each of the necessary conditions to establish 'a self-evident and significant proposition'. Indeed, when we examine the previously stated analyses and the following arguments, we will notice that Sidgwick is going to establish ethical hedonism 'in the highest degree of certainty attainable', by (1) clearly defining the terms 'pleasure' and 'good', (2) ascertaining that the claim of ethical hedonism is nontautological, (3) confirming its truth by careful reflection, (4) examining whether there is a general consensus about it, and (5) verifying that ethical hedonism can coherently systematize our actions.

7.2.4.1 Step one: Argument of elimination

The ultimate good is limited to what relates to human consciousness. Sidgwick's proof of hedonism begins with the process of narrowing the scope of the ultimate good. The point to be recalled here is what was suggested at the end of Sidgwick's analysis of the concept of good. Sidgwick stated there that 'if we consider carefully such permanent results as are commonly judged to be good, other than qualities of human beings, we can find nothing that, on reflection, appears to possess this quality of goodness out of relation to human existence, or at least to some consciousness or feeling' (*ME* p. 113; 4.5 of the present book).

Good considered here is 'Good [. . .] to be sought by man, as an ultimate practical end' (*ME* p. 115). This comes as no surprise, since the good mentioned in the Principles of Self-love and Benevolence is the good that a person should pursue. If we, in accordance with this assumption, are to consider a good which might become the ends of

our action, we will first notice that someone must be conscious of such a good. Even if there is some good at a distance that no one can yet perceive, that is not relevant to ethics, which deals with how to decide our voluntary actions.

'Things that a person can be conscious of' include almost everything that we can think of, for example, lifeless objects such as money and works of art, or intangibles such as knowledge and ideals. However, Sidgwick points out that, though lifeless objects can exist apart from our consciousness, we judge them to be good so long as they come into our consciousness. We also judge beautiful works or lofty ideals only when we are aware of them. If this is correct, presumably those lifeless objects or ideals are good *by virtue of* their connection with human consciousness, and not good in themselves. Thus Sidgwick limits the scope of ultimate good to something inseparable from human consciousness.

Of the things inseparable from human consciousness, what we call good or bad would be either consciousness itself, or human existence, action or character that always involves consciousness. Among the various types of consciousness, the most promising candidate for the ultimate good would be, of course, 'desirable feelings', that is, happiness or pleasure. Goodness in human character or action is often described as 'perfection' or 'excellence'. On this account Sidgwick maintains that 'if there be any Good other than Happiness to be sought by man, as an ultimate practical end, it can only be the Goodness, Perfection, or Excellence of Human Existence' (*ME* p. 115).

To proceed a little further, when we talk about the goodness of human existence, some people say that human existence or human life is good for its own sake. We also think that the main component of human perfection or excellence is virtue (see 3.1.4 of the present book). Some people insist that, though virtue is perceived through excellence in character or action, its essence consists in 'good will', namely the subjective goodness of the agent's will. Will is also a kind of consciousness, and therefore can be a candidate for the ultimate good. In the previous paragraph we mentioned only happiness as a candidate for the ultimate good; but other types of consciousness may be regarded as the ultimate good.

Thus, one by one, Sidgwick examines these candidates for the ultimate good. They include virtue as excellence in human character and action, the subjective goodness of will, and human existence – all of them being inseparable from human consciousness. Sidgwick rejects most of them, and finally concludes that 'desirable feeling' is the sole candidate that can survive his examination.

Virtue. Sidgwick first examines whether virtue is the ultimate good (Bk. 3 Ch. 14 Sec. 1). As explained in 3.1.3 of the present book, virtue is an excellent quality that some humans have, which is relatively permanent and can be fostered by one's voluntary effort. It is commonly believed that such a quality is manifested as excellence in character or action. According to Sidgwick, however, we cannot directly evaluate one's character. A person's abilities or disposition that constitute his character can only be defined as the tendency to *act* or *feel* in a certain way under certain conditions. When we judge someone's character to be virtuous, our judgment actually depends on our evaluation of the actions or sentiments that he consistently displays, or of the consequences of such actions or sentiments (*ME* p. 393). This being so, the plausible candidate for ultimate good would not be the quality of a person's *character* but the quality of his *actions* or *sentiments*.

When we examine virtue as something manifested in actions and sentiments, virtue is called a human quality manifested in right actions (duties) and other good actions or sentiments. Common-sense morality specifies such actions as particular duties and virtues. However, if we define 'being virtuous' as 'generally observing common duties and virtues', obviously we cannot regard it as the ultimate good, because this is a circular argument. In 5.3 and 7.1.1 of this book, we have concluded that common moral rules are very often ambiguous, and that in order to precisely stipulate them we need higher principles, that is, the three fundamental principles. Next we found (in 6.1.2, 6.1.3, 7.1.2 of this book) that two of the three principles involve the concept of ultimate good, and now we are trying to clarify the content of this ultimate good. If we insist that the ultimate good is to observe common moral rules, this will become an empty argument that goes round and gets nowhere (*ME* p. 392).

Sidgwick further claims that particular virtues of common-sense morality, such as zeal, courage, temperance, wisdom, benevolence and justice, cannot be regarded as the ultimate good – though we may say that they constitute *part of* the ultimate good, if we take a certain viewpoint. I will return to this later.

First, virtues such as zeal or energy are praiseworthy only when they are manifested in pursuing good ends. Their goodness is judged by some other criteria than whether the action was zealous, energetic, etc. or not. This means that they are good only as a means to other forms of goodness.

Second, some common moral virtues are called good only within a certain limit; and when we try to precisely determine where this limit

lies, we notice that we need to refer to other maxims, good, or ends, such as happiness. For example, generosity, courage and frugality become extravagance, foolhardiness, and meanness when they exceed a certain limit. Then, how can we discern the line between virtue and excessive behavior? Not by dogmatic intuition that dictates those virtues, but probably by appealing to some higher axioms, or to some other good, such as happiness. Obviously, when those behaviors bring about great pain to others, or when they prevent the agent from attaining a higher good, they are not regarded as virtues. This also suggests that these kinds of common virtues are not good in themselves.

Certainly, there are other virtues that seem to have no such limits. Three well-known virtues – wisdom, benevolence and justice – are usually regarded as having no limits on goodness. However, careful analyses of these virtues show that the precise expression of these virtues inevitably involves the concept of good. Wisdom is an ability to discern a certain good and the means to that good; benevolence is manifested in the action of doing good to others; and the aim of justice is to fairly distribute good things (advantages) or bad things (burdens) according to a certain rule – and we cannot find any better explanation for these virtues. This being so, it becomes a vicious circle to answer, when asked what good is, that it is the wisdom to discern good, benevolence to do good to others, or the impartial distribution of good and bad things. These are empty answers. Since these virtues presuppose the concept of good, it is pointless to consider them to be the ultimate good. Thus, any one of these particular virtues of common-sense morality cannot be regarded as the ultimate good.

The subjective rightness or goodness of will. Sidgwick next examines the subjective goodness or rightness of will, which is often said to constitute the essence of virtue (*ME* Bk. 3 Ch. 14 Sec. 2). This argument was added to his proof of hedonism in the fourth edition of *ME*. What Sidgwick has in mind here is, undoubtedly, Kant's argument on good will. Will which is subjectively good or right means a will to perform what one judges to be right, and to bring about what one judges to be the best. 'What one judges to be right' and 'what one judges to be the best' can be *whatever* one judges to be right or best, and they do not presuppose the existence of some objectively established good (that is, good which is discernible to others). Therefore, we cannot reject good will as a candidate for the ultimate good on the grounds that it presupposes another good. However, Sidgwick claims that, inasmuch as good will has nothing to do with any objective good, it cannot be the ultimate good.

We should first note that we are currently concerned not with subjectively right actions but with objectively right ones, that is, actions that are judged to be right for reasons *other than* that the agent *believes* in their rightness (1.2.4 of this book). We need a concept of ultimate good that enables us to make judgments on the objective rightness of actions. Even if we conclude that good will is the ultimate good to be sought, this will only urge us to perform *whatever* is an objectively right action, and it never helps us judge what actions are objectively right. It is paradoxical to tell a person that it is good to seek to perform the right action, when he is trying to find out what good makes an action right.

In addition, we sometimes make judgments on the rightness or wrongness of others' acts, e.g. 'his action is objectively wrong, though he certainly is doing what he believes to be right'. Our judgment that his action is objectively wrong is not based on the agent's will but on the ulterior bad consequences of his action – in most cases, the effect of such action on people's happiness or pain. If this judgment makes sense, then the ultimate good that is needed for judging the objective rightness of actions should not be the agent's good will but the other good that is found in those ulterior consequences.

There is another argument which shows that we cannot regard the subjective goodness of will as the *sole* ultimate good, which is as follows. We do not think it desirable to always consciously have good will. Rather, we think that things often go well when we act from other motives, without being conscious of the will to perform the right action. Sidgwick does not give examples of this, but a manufacturer of airplanes or a craftsman of wheelchairs, for example, may perhaps be able to create his best products when he puts aside his conscious will to help others by creating those products and simply concentrates on the unfinished work before him. If this is true, we think that it would be more desirable for such a manufacturer or craftsman to put aside his good will for a while. Therefore, it is meaningful to ask whether, and to what extent, we should consciously seek to have subjective good will. This suggests that the subjective goodness of will is not the only ultimate good to be reasonably sought.

Sidgwick previously stated that 'if there be any Good other than Happiness to be sought by man, as an ultimate practical end, it can only be the Goodness, Perfection, or Excellence of Human Existence'. However, it now turns out that virtue, which is a component of perfection or excellence, cannot be identified with the ultimate good, as long as we see it as (1) a general obedience to the rules of duty and virtue, as (2) particular components of virtue, or as (3) the subjective goodness

of will that has nothing to do with any objective good. The possibility remains, however, that virtue could be part of the ultimate good when we look at its other aspects. According to Sidgwick, neither can other types of excellence, such as special talents or skills, be called the ultimate good. This is because those talents or skills are regarded as excellent only when they produce some other good or promote happiness (*ME* p. 395).

Human existence. Thus we have to look for other candidates than excellence or perfection in human beings. Then, how about human existence itself? Can't we consider human life itself to be intrinsically good?

Regarding this, Sidgwick argues that we should clarify exactly which aspect of human life we think to be truly desirable. Then he points out that, when we exclusively consider the physical aspect of our existence – or when we regard our existence merely as complex processes of physical change or as complex movements of particles of organized substance – we cannot seriously think such physical existence is good in itself regardless of whether it involves consciousness or not. I suppose most of us will admit this point. If a member of our family fell into a vegetative state in which he is physically functioning without consciousness, we would certainly regard it as a very bad situation. We regard a person's existence as good probably because it usually involves consciousness. If this is correct, a person's physical existence is not desirable in itself. Rather, the ultimate good seems to be something connected to the conscious aspect of a person's life, or to consciousness itself. Hence Sidgwick states that 'In short, if a certain quality of human Life is that which is ultimately desirable, it must belong to human Life regarded on its psychical side, or, briefly, Consciousness' (*ME* Bk. 3 Ch. 14 Sec. 3 p. 396). When we value our conscious life this way, several ex-candidates for the ultimate good, such as virtues and good will, might be reevaluated as constituting *part* of the ultimate good. *To act* upon duties and virtues, or *to will* to realize virtues are part of a conscious life, and as such they may constitute part of the ultimate good.

Consciousness. However, a 'conscious life' or 'consciousness' cannot be unconditionally regarded as good. A conscious life involves not only pleasure but also pain, and a life full of pain is usually considered to be undesirable. There are also other types of undesirable conscious life – boredom, monotony, emptiness, etc. (The question of what makes a life 'undesirable' is irrelevant. What Sidgwick describes here is a consciousness that contains nothing we judge to be desirable – no happiness, no

liberty, no beauty, or whatever.) If such an undesirable consciousness lasts throughout one's life, Sidgwick claims, we would not think it good just to prolong such a life. Most of us would admit this point, too. When we take the euthanasia controversy seriously, we do not just discuss whether a patient has consciousness but whether his consciousness is a desirable one or not. Certainly, we generally respect life and make an effort to save and prolong one's own and others' lives; we generally think it wrong to destroy lives; and we think the most important thing is to foster social habits and moral sentiments to maintain human lives. But all these are not because the mere existence of a conscious organism is desirable but because we assume that our lives normally contain consciousness that is by and large desirable.

Similarly, if we judge virtuous activities as constituting part of the ultimate good, it is not simply because they are part of conscious life but because the consciousness which accompanies such activities is normally regarded as intrinsically desirable by the agent (*ME* p. 397). Thus, Sidgwick argues that the most promising candidate for the ultimate good is '*desirable* consciousness' (*ME* p. 397).

Cognition, volition and pleasure. At this point, we might be tempted to immediately conclude that this 'desirable consciousness' is pleasure or happiness, and that therefore pleasure is the ultimate good. Certainly, we generally regard life to be good, assuming that our lives are normally *happy* on the whole. No one would claim that the consciousness of a virtuous martyr while undergoing extremely painful torture is in itself desirable – though it might be his duty to receive pain for the sake of someone else's good, or it might be in his own interest to feel pain at present because that enables him to obtain the ultimate happiness he seeks.

Sidgwick is cautious, however. According to him, our conscious experiences include not only *feeling* but also *cognition* and *volition*; and some may claim that the desirability of the latter two aspects of consciousness should not be measured by the same criterion as that for pleasure, that is, the criterion based on the desirability judged by the experiencing individual at the time of experiencing it. Indeed, certain consciousness other than feeling can sometimes be judged to be desirable. For example, knowledge of truth or contemplation of beauty are *cognitive* states which are often judged to be desirable. A will to perform a right action, or a will to realize virtue or other social ideals (say, freedom) are *volitional* states, which are often judged to be desirable. In order to argue that pleasure is the sole ultimate good, we need to show that cognition

and volition are not desirable in themselves, and that what is intrinsically desirable is feeling, and feeling alone.

Then Sidgwick argues that *cognition* and *volition* are not desirable in themselves, if considered separately from their consequences, the evaluation of truth/ideals, and the accompanying feelings. When we solely consider one's mental experiences of cognition or volition, they are utterly neutral in terms of desirability.

Indeed, knowing truth, contemplating beauty, and having the will to attain a certain ideal are sometimes considered to be more desirable than feeling pleasure. Sidgwick believes, however, that when we claim this, we are confusing the value judgment of a *conscious state* (such as cognition or volition) with the value judgment of the *nonconscious state* that accompanies such cognition or volition. In such a case, what we actually judge to be good or desirable is either (1) future effects which are expected to be produced by the current consciousness, or (2) a certain objective relationship between the conscious subject and the external things that exist apart from his consciousness.

> It is no doubt true that in ordinary thought certain states of consciousness – such as Cognition of Truth, Contemplation of Beauty, Volition to realise Freedom or Virtue – are sometimes judged to be preferable on other grounds than their pleasantness: but the general explanation of this seems to be (as was suggested in Book ii. chap. ii. § 2) that what in such cases we really prefer is not the present consciousness itself, but either effects on future consciousness more or less distinctly foreseen, or else something in the objective relations of the conscious being, not strictly included in his present consciousness.
>
> (*ME* p. 399)

We will need to further explain these 'objective relations'. Let us take the example of knowing truth, which is often regarded as preferable to feeling pleasure. In order that a person can truly say that he knows a certain truth, it is not sufficient for him to *believe*, within his consciousness, *that* he knows the truth; rather, he must be recognizing a real state of affairs that exists *outside* his consciousness. When we judge the knowledge of truth to be desirable, we are evaluating the formation of this relationship of 'to know/to be known' between his internal consciousness and the external thing. This relationship can be called 'objective' because it relates to a state outside his consciousness. Similarly, when we regard the volition to fulfill virtue or the contemplation of beauty as desirable, we are not simply evaluating one's will to perform

what *one believes* is virtuous, or one's contemplation of what *one believes* is beautiful; rather, our evaluation is based on the assumption that one's view of virtue or beauty meets certain objective criteria for true virtue and beauty. In other words, we regard the volition to fulfill virtue or the contemplation of beauty as desirable *only when* an 'objective' connection holds between one's consciousness (i.e. thought about beauty or virtue) and a certain ideal form of beauty or virtue, which is supposed to exist independently from one's consciousness. Again, when we judge the life of a poor free man to be preferable to that of a rich slave, we are claiming not that one's *internal cognition* of freedom is desirable in itself, nor that one's *internal determination* to be free is desirable; rather, we are claiming that it is desirable for a person to be free from another's will. Here, we are evaluating the formation of an objective relationship between one's consciousness and nonintervention from others. In all the cases just described, the true object of our value judgment is not consciousness itself but a certain 'objective' relationship between one's consciousness and something outside of it. Evidence of this can be given by the fact that, if it turns out that what a person believed to be truth, beauty, liberty, etc. is after all illusory, we will no longer regard the consciousness he once had (i.e., the deluded consciousness that 'I know the truth', 'I am witnessing beauty', 'I am free', etc.) as intrinsically desirable.

Thus Sidgwick contends that, in considering a certain cognition or volition as desirable, what we actually judge to be desirable is either (1) its future effects or (2) a certain objective relationship. In such cases, we are not evaluating the *consciousness* of cognition or volition itself, since neither (1) nor (2) is consciousness. Meanwhile, in our previous arguments we have already narrowed the candidates for the ultimate good to 'desirable consciousness'; so both (1) and (2) are already excluded as candidates, since they are not consciousness. Thus, of the three aspects of 'desirable consciousness', two, 'desirable cognition' and 'desirable volition', are eliminated as candidates for the ultimate good. The only remaining candidate is 'desirable feeling', that is, pleasure.

As for pleasure, we have already concluded that we are evaluating the pleasant feeling itself when we call it desirable. Even if a pleasant feeling turned out to have been caused by an illusion, the person who experienced it can still think that the feeling itself was certainly desirable, provided that he considers it separately from all other factors that accompany that feeling.

Thus, the consciousness that we can truly call desirable is pleasure, and pleasure alone. At the same time, there remains no other candidate

for the ultimate good. Therefore, ethical hedonism is the only one we can support as a plausible theory for the ultimate good.

With the above arguments Sidgwick's proof of hedonism is completed for the most part. However, Sidgwick admits that this conclusion might not have been derived if we had defined 'a conscious life' in a different way. So far we have denied the possibility that cognition and volition be called desirable, on the grounds that the 'objective relationship between a conscious subject and an external state' (which is implied by the concepts of virtue, truth, beauty, liberty, etc.) cannot be included in one's consciousness. However, if we interpreted 'one's conscious life' in a much wider sense and included such an objective relationship in it – for this relationship is undoubtedly an indispensable component of a conscious being's cognition or volition – we might have concluded that knowledge of truth, contemplation of beauty, volition to produce virtue, etc. can be rightly called 'desirable consciousness', and that such 'desirable cognition/volition' can also be regarded as the ultimate good. Then we might have even claimed that they are often preferable to pleasure.

Admitting this possible objection to hedonism, Sidgwick endeavors to verify his proof of hedonism by appealing to an individual's reflective intuition and to people's commonsense (*ME* Bk. 3 Ch. 14 Sec. 5 p. 400 ff.). He attempts to show that, according to our reflective intuition and commonsense, (1) the 'objective relationship' described above is always evaluated by its contribution to the promotion of pleasure, and that (2) happiness is always explicitly or implicitly assumed to be the ultimate good.

7.2.4.2 Step two: Appeal to intuition and commonsense, and replies to objections

First, let us look into the argument in which Sidgwick appeals to each individual's reflective intuition. He admits that we often judge something other than consciousness to be ultimately desirable. However, he asserts that, when we carefully ask ourselves why such qualities as truth, beauty and virtue are so important, we cannot find any other reason than that they contribute to the happiness of sentient beings (*ME* p. 401).

Next Sidgwick examines our ordinary judgments, or the judgments of our commonsense. At a first glance, our commonsense seems to disapprove of ethical hedonism. We cannot deny that some cultured people often judge that we ought to pursue knowledge, beauty, virtue and other ideal 'goods' independently from the pleasure that can be obtained from them. Sidgwick points out, however, that in fact such

ideal goods tend to produce pleasure in various ways. It even seems that the more pleasure they produce, the more our commonsense regards them as desirable; and we seldom claim that they are desirable even when they have no tendency to promote happiness. Obviously, beauty and social ideals will not be regarded as desirable when they have no tendency to produce any happiness. Knowledge, too, seems to be evaluated higher when it brings about more fruitful results. It is often claimed that it is also desirable to explore academic knowledge even if it does not seem to produce noticeable benefits; but this is probably because (1) knowledge often bears unexpected results that provide us pleasure, because (2) such knowledge may shed light on other areas of study that are seemingly unrelated to it, because (3) it brings the inquirer the pleasure of satisfying one's intellectual curiosity, or because (4) one's intellectual disposition, which is trained by continuous study, is likely to produce fruitful knowledge in the long run. At the same time, commonsense does not always praise the pursuit of knowledge. Whether it is desirable or not seems to be determined by the effects of such a pursuit on the happiness of human or sentient beings. To take Sidgwick's example, when we discuss whether it is desirable to obtain knowledge about the human body by performing vivisection, both pros and cons overtly appeal to its effects on happiness to justify their own claims. In today's context, the debates over research on human cloning would pose a similar argument.

To take another example of pursuing virtue, there are cases in which a person's efforts to improve his virtue become so fanatical that he would even sacrifice his own and other people's happiness. In such cases, we commonly question whether such an effort is really desirable. Especially when such moral fanaticism actually brings, or is expected to bring, unhappiness to people, we tend to measure the depth of the expected unhappiness before determining its desirability.

Thus, according to our common-sense judgments, when we carefully reflect on seemingly desirable activities such as knowing truth and contemplating beauty – and especially when a question or dispute arises regarding their desirability –, we naturally adopt pleasure as an criterion to determine the desirability of such activities.[10]

Still, many ordinary people seem to have a general antipathy toward the view that pleasure is the sole ultimate good and the ultimate end of right actions. Sidgwick thinks, however, that such an antagonism is based on misunderstanding, and will disappear if such misunderstanding is cleared up. (*ME* Bk. 3 Ch. 14 Sec. 5. Sidgwick divides his arguments into four main points, but I will summarize them in three points.)

First, in our ordinary usage, the term 'pleasure' is used in a very narrow sense, and it is most often used to denote primitive kinds of enjoyment. However, Sidgwick uses this term in a much wider sense, to mean any feeling that is felt as desirable. We sometimes regard certain pleasures as undesirable, but that is because such pleasures are expected to cause more pain than pleasure. What we judge to be undesirable in such cases is not the pleasure itself but the pain (or loss of pleasure) that is caused by such pleasure.

Second, hedonism is quite often equated with egoistic hedonism. Therefore, people misunderstand the claim that 'happiness is a person's sole ultimate good' as insisting that 'one ought to pursue *one's own* happiness even if it sacrifices others'; then they oppose hedonism on the grounds that they abhor the latter claim. However, what Sidgwick has proved is *ethical hedonism in general*, that is, the claim that the basic element of good is pleasure, whether it is one's own good or other people's good. This *ethical hedonism in general* does not exclusively support egoistic hedonism, so this type of objection is essentially irrelevant.

Third, whether it is a person's own happiness or people's happiness, pleasure is often better attained when we do not consciously pursue it. This is because we can act more effectively when we concentrate on a definite purpose, rather than when we foresee, measure and calculate a particular pleasure that will result from our conduct. In addition, when we are pursuing a certain ideal, we will not be swayed by contingent factors and will eventually succeed in attaining happiness. If these are empirically admitted facts, even from the viewpoint of hedonism, the pursuit of such a definite purpose or ideal can be indirectly regarded as rational. Therefore, it is no wonder that commonsense highly evaluates certain purposes and ideals. Nevertheless, when we have to determine how far we should pursue such purposes or ideals, or which to pursue when we are inclined to pursue two mutually exclusive purposes, we will consider the degree to which such a pursuit will contribute to happiness. If this is true, our commonsense implicitly admits ethical hedonism.

Thus Sidgwick claims that there is no reason to think that hedonism runs against commonsense.

7.2.4.3 Step three: Argument from practical need

We cannot say, however, that the foregoing observation on commonsense *fully* showed that there is 'universal consensus' or 'consensus among experts' about the truth of hedonism. Sidgwick himself seems

to be concerned about the possibility of his argument being denied (*ME* p. 406). Thus, he finally emphasizes that we cannot find any other theory that would better meet our practical needs.

> If we are not to systematise human activities by taking Universal Happiness as their common end, on what other principles are we to systematise them?
>
> (*ME* p. 406)

If there is a systematic theory of good, it should be one that enables us to compare various good things. We need to decide what we ought to now pursue, among different things that we usually regard as good. When we cannot simultaneously pursue various ends, such as beauty, truth, ideals and faith, we have to decide which of them to seek. In order for us to decide, we have to compare the various 'goods'.

Happiness will be certainly included in that array of goods, for it is paradoxical to regard desirable feeling as *not* desirable. Therefore, when we foresee someone's pain (or loss of pleasure) as the result of pursuing a certain good, we have to balance that good with the accompanying pleasure/pain and decide how far we ought to pursue that good. Thus, if there is a theory of good, it should also provide a common standard to compare pleasure with other goods.

Hedonism is the very theory that affords such a standard. It is commonly believed that pleasure can, at least for practical purposes, be measured and compared with sufficient clarity. Moreover, as for pleasure we commonly accept the criterion that the quantity of pleasure is to be measured and compared in terms of 'its desirability from the viewpoint of the experiencing individual' – though we also admit that its *strict* measurement is difficult in practice. By adopting this common criterion of comparing *pleasures* as the weighing of various *goods,* we can compare 'goods' by assessing the pleasures that they generate.

Then, is there any other theory that enables us to compare different goods more clearly than hedonism? Sidgwick is convinced that such a theory cannot be found. If this is correct, ethical hedonism, or hedonism regarding the ultimate good, is the only theory that meets our practical needs.

Thus, ethical hedonism, or hedonism concerning the ultimate good, gets support from these three-layered arguments: (1) pleasure is the only candidate for the ultimate good that survives philosophical analysis, (2) pleasure is also what our reflective commonsense is likely to regard as the ultimate good, and (3) hedonism is the only theory

that can meet our practical needs. In short, ethical hedonism is the only alternative left as a plausible theory on the ultimate end to be pursued.

7.3 Basis of utilitarianism

According to Sidgwick, the ultimate good is pleasure. When this proof of hedonism is combined with the maximization principle obtained at the end of 7.1 of this book, the utilitarian principle is finally established. In other words, when ethical hedonism is integrated into the principle of 'equal consideration for each individual's good at every point in time, and the maximization of the sum total of those goods', we can derive the principle of 'equal concern for people's *pleasures* at every point in time, and the maximization of the sum total of those *pleasures*'. This final principle is the very claim of universal hedonism, that is, utilitarianism.

Let us look into this derivation in more detail. According to our conceptual analysis of good, *one ought to act so as to bring about the greatest attainable good on the whole*, provided that one will choose the right action by considering the good to be pursued (in other words, provided that one adopts a teleological or consequentialist position about what one ought to do). 'Good on the whole' mentioned here is what one now judges to be good after considering all the 'goods' of various people. The Principles of Rational Self-Love and Benevolence impose certain restraints on our judgment of this good on the whole. In considering such good on the whole, these two principles require us to *equally* treat different 'goods' of each individual at each point in time *in proportion to their magnitude*. What we judge to be 'the good on the whole' after fairly weighing and aggregating different 'goods' would be expressed as the *sum total* of 'goods' of all parties concerned at all points in time. Thus, *as long as* we accept the Principle of Rational Benevolence and take an impartial point of view, and *as long as* we accept the Principle of Rational Self-Love in considering the good on the whole for any individual, *we ought to equally treat the different goods of each individual at each point in time in proportion to their size, and to act so as to maximize the sum total of those goods*.

Before ethical hedonism was proved, however, the criterion to compare the magnitude of different 'goods' had not been determined. This criterion must settle which good ought to be pursued. It must also be a universal criterion accepted by all rational beings. Now that ethical hedonism is proved, we have such a criterion. The criterion for

comparing goods is pleasure. The greatness of good is measured as the greatness of pleasure accompanying the good in question. This judgment on the greatness of pleasure is based on how strongly it would be desired by the individual at the time of feeling it, when he exclusively considers that feeling apart from any other factors. Sidgwick presupposes that the individual experiencing pleasure can recognize its greatness, or the degree of its desirability, and determine its value. On this assumption, we can compare the magnitude of pleasures for each individual at each point in time, and meaningfully call one pleasure that has the same value as another pleasure (albeit that of a different individual at a different time) 'a pleasure as great as another'.

If pleasure is the sole ultimate good, and if we can think of no other feasible method of comparing various 'goods' than the hedonistic one, the most plausible explanation about the relationship between good and pleasure would be as follows. First, an individual's ultimate *good* at a certain point in time (i.e. what he ultimately judges to be desirable when he exclusively considers himself) would be the *pleasure* that he feels at that point (i.e. the feeling that is desirable for him at that point). Second, the *magnitude of the ultimate good* for each individual at each point in time would correspond to the *greatness of pleasure* for him at that point; to use different ways of determining its magnitude would be to introduce criteria other than pleasure. Third, if we may assume that we can compare the size of pleasures of various individuals at different times, and that we can meaningfully talk about 'a pleasure as great as another', the phrase 'a person's *good* which is as great as another's good at a different time' would correspond to the phrase 'a person's *pleasure* which is as great as another's pleasure at a different time'. Thus, if we adopt consequentialism, ethical hedonism and the Principles of Rational Self-Love and Benevolence, *plus* if we are ready to take an impartial viewpoint as the Principle of Benevolence suggests, *we ought to equally treat every individual's pleasure in proportion to its greatness, and to maximize people's pleasure as a whole*. We have already suggested that this measuring of pleasure can be expressed as the amount or quantity of pleasure; therefore, 'to equally treat people's pleasures in proportion to their greatness' is to equally treat them in proportion to their amount. To maximize them as a whole would be to maximize the aggregation of people's pleasures, and the simplest way to aggregate the amount of pleasures is to sum them up – in short, the aggregation of people's pleasures would be expressed as their *sum total*. Thus, finally, we have obtained the typical utilitarian principle of 'the maximization of the sum total of people's pleasures'.

7.4 Utilitarianism and commonsense

The theoretical basis of utilitarianism has thus been ascertained. Sidgwick further argues that our reflective commonsense also supports utilitarianism.

Certainly, when we make moral judgments, we are not always guided by conscious inferences, and we do not always foresee the general happiness which may result from our actions. We often act morally simply out of a sense of duty, or according to our moral intuition. Nevertheless, our common-sense morality can be explained in utilitarian terms, as follows:

(1) First, utilitarianism can clearly explain the general validity of commonly accepted moral rules. For example, it is evident that the habit of keeping a promise, telling the truth, etc. generally contributes to people's happiness, for it generates and maintains mutual trust and enables us to enjoy various activities that are fostered by a cooperative relationship. Some critics object that utilitarianism cannot explain one's special duty to close companions, such as a parent's duty to his or her children. However, utilitarians can explain why each person should foster a special affection to people close to him, instead of always caring for humankind in general. We would be able to maximize people's overall pleasure if we primarily care for the happiness of people with whom we are familiar, because we are in the best position to recognize such people's needs and desires. Other critics may claim that utilitarians cannot explain the virtue of temperance, which tells us to abstain from having pleasure. However, the pleasures that this virtue condemns are, mainly, those kinds which have the tendency to damage one's health, or which tend to obstruct the development of other capacities or sensitivities that would become the source of happiness (*ME* p. 449). When we enumerate the qualities of action and character that would directly or indirectly produce people's happiness, they seem to include all the common virtues and duties (see, for example, *ME* p. 424).

(2) Second, utilitarianism can also explain the widely recognized limitations and exceptions to common moral rules. For example, it is often believed that we do not have to fulfill the duty of keeping a promise when (i) the promisee nullified the promise, when (ii) the promise was extracted through fraud or coercion, when (iii) the situation has so greatly changed that the result of keeping a promise is likely to be quite different from what was expected at the time the promise was made, or when (iv) keeping the promise will bring about great harm to the promisor or the promisee. Utilitarians would rightly admit cases (i)–(iv) as exceptions to

the duty of promise-keeping, because in cases (i) and (iii) neither promisor nor promisee will gain the happiness that was expected to result from the promisor keeping the promise; because it would be disadvantageous to a society if we admit the validity of fraud or coercion in such cases as (ii); and because keeping a promise in case (iv) would obviously damage people's happiness.

(3) Third, when we discuss the proper definition and appropriate scope of a moral rule, most of us naturally refer to this rule's effects on people's happiness. To take the example I mentioned in my discussion of the proof of hedonism, wisdom cannot be defined except as the rule that 'one ought to choose the best means to the right end', and when we ask what this 'right end' is, most of us would uphold the goal of someone's happiness. Additionally, when we are divided as to the interpretation of a certain rule, each side of the debate usually supports its own claim by arguing that this rule, so interpreted, will eventually contribute to people's happiness. For instance, we may debate over the question of whether the rule that 'one ought to show gratitude to one's benefactor' implies that the requital should be based on the benefactor's *effort*, or on the *result* that the benefactor actually brought about. In another example, we may discuss how far we should pursue scientific research. In such cases, both sides of the debate generally appeal to a consideration of the ultimate effect of such rules or research on people's happiness.

(4) Furthermore, when two or more moral rules come into conflict, most of us seem to arbitrate it by ultimately appealing to the effect of our decision on the general happiness. For example, when a parent's duty to his or her child conflicts with one's duty as a professional (as when a doctor has to choose between fulfilling a promise to bring her child to Disneyland and carrying out her duty at the hospital on the same day), we seem to ultimately decide by considering which course of action will bring about more happiness, or less unhappiness, in that particular situation. Likewise, when we are debating which distribution principle, that according to merit or that of equal distribution, is better in a particular situation, we seem to weigh the benefits of encouraging people to engage in meritorious actions, which would promote general happiness, against the negative effects of envy which may result from unequal distribution.

(5) Utilitarianism can also explain the reasons why people's duties and virtues often vary according to their occupations and positions, and the reasons why moral rules often differ from nation to nation and from time to time. For example, we often tolerate a certain degree

of dishonesty in diplomats or salesmen, such as not readily admitting the defects of their policies or products, while we seldom allow court justices to conceal the truth. Such a variance can be explained by utilitarians, because it corresponds to the different ways in which various occupations affect people's happiness. Similarly, the diversity of moral rules at various times and among nations corresponds to the differences in the effects, or in the expectation of the effects, of those rules on people's happiness. For example, lending money is condemned in a society where commerce is underdeveloped, and vengeance is often tolerated in a society whose penal system is inadequate (*ME* pp. 453–4).

(6) We must admit, however, that our moral sentiments are often in discord with utilitarian precepts; but we can coherently explain why this happens (*ME* pp. 455–6). Even if a certain moral rule was originally laid down through utilitarian consideration, our current moral sentiments may dominate our ordinary moral thinking. This naturally happens, and is a good thing even from a utilitarian point of view, for it is quite useful for us to foster and maintain strong moral sentiments that urge us to obey rules in appropriate situations, and even during emergencies. Nevertheless, such sentiments are sometimes overemphasized, and by erroneous analogy we may come to hold the same sentiments toward seemingly similar but irrelevant situations. Furthermore, such sentiments, being ingrained in people's minds, may survive even after the original moral rule became useless from the utilitarian viewpoint, due to the changes in time and situation. This is why we often feel as if our moral sentiments run against the utilitarian judgment in a new situation. However, if we reflect on our own sentiments and judgments, we start to question why we have to observe a moral rule which no longer contributes to the general happiness; and we are often led to adopt a new moral rule.

Thus, Sidgwick proposes a supposition that the general validity of common-sense morality is 'unconsciously' supported by utilitarianism. According to him, we are not always conscious of the effects of our decision on people's happiness in deciding what we ought to do; but when we pause to reflect, we will find that the ultimate rationale of our moral judgments is a utilitarian one. Sidgwick calls this the hypothesis of unconscious utilitarianism[11] (*ME* p. 454. See also *ME* pp. 401–2, 424, 450, 453, 456, 463). If this hypothesis works, utilitarianism is also supported by common-sense morality.

Then, why do we intuitively follow common moral rules, instead of always consciously making a utilitarian calculation? We can read the following points in Sidgwick's arguments.

The most obvious method of utilitarian thinking is to adopt empirical hedonism (*ME* Bk. 4 Ch. 4 p. 460), which first ascertains the alternatives of action that are open to us in a particular situation, then considers foreseeable pleasures and pains that could result from each alternative, compares the alternatives and finally chooses one that will bring about the greatest overall happiness. This method involves certain difficulties and uncertainties. There is every likelihood that one might make a mistake in forecasting one's own pleasure, and the difficulty gets even greater when one has to consider the effects of one's action on all affected parties. Probably, the rules of common-sense morality were devised in order for us to deal with such difficulties. In other words, such rules can be regarded as the expressions of 'positive beliefs of mankind as to the effects of actions on their happiness' (*ME* p. 461). Having undergone a long period of refinement, common moral rules are likely to be made so as to properly cope with complicated emergencies in most cases (*ME* pp. 475–6). From the utilitarian point of view, it is generally useful for people to habitually abide by such rules with strong moral sentiments. So considered, we can say that common moral rules can serve as the 'middle axioms', whose validity is supported by the utilitarian principle, and which guide us to perform certain actions in particular situations.

However, such middle axioms do not necessarily make up a coherent system, as we have seen in our examination of common-sense morality. Additionally, the established duties and virtues, which have been fostered through history, do not always maximize people's happiness for several reasons (*ME* pp. 463–7). Such rules may be more or less distorted by people's insufficient knowledge of the causal relationships between action and consequences, by the limits of people's imagination of other's pleasure and pain, or by the influences of authority and customs. These rules may no longer be suitable for today's context, as people's desires or circumstances might have changed. Common moral rules are only applicable to ordinary people in ordinary situations, and not intended to deal with an extraordinary or rare one. Thus, we often have to reconsider whether a certain rule should be applied to *this* particular individual in *this* particular situation. Moreover, as we have already seen, when we are questioning the precise definition or the proper scope of a certain rule, or when it conflicts with other duties, we cannot coherently determine what we ought to do just by adhering to common duties and virtues. Therefore, we have to guide our actions systematically by consciously following utilitarian thinking, when needed, to supplement the inadequacies of our common-sense morality, while generally respecting our common moral rules.

7.5 Utilitarianism and egoism

Utilitarianism thus has solid theoretical foundations as well as sufficient support from common-sense morality. However, we should note that a serious problem arises here. Sidgwick's proof of hedonism, examined in the previous section of this book, was the proof of *ethical hedonism in general*, which is the common basis for two types of hedonism, egoistic hedonism and universalistic hedonism. When this proof of hedonism is incorporated only into the Principle of Rational Self-Love, which concerns one's own good, we can obtain another principle, which is that 'I ought to perform an act that will maximize my pleasure on the whole'. As a result, egoistic hedonism can be as firmly endorsed as utilitarianism.

Let me explain this in more detail. When I solely consider my own good on the whole, I am required, by the combination of ethical hedonism and the Principle of Rational Self-Love, to *equally treat* my present pleasure and my future pleasure according to their magnitude. Then, *provided that I exclusively attach importance to my own good*, I am led, by the definition of the concept of good, to endorse egoistic hedonism, that is, the claim that I ought to perform an action which will maximize the sum total of my pleasure.

It might be objected that, since everyone should also admit another self-evident intuitive principle, that is, the Principle of Rational Benevolence, we should also treat others' good as equally important to one's own, and therefore are led to accept utilitarianism. However, here we have to recall the meaning of the phrase 'everyone admits the Principle of Rational Benevolence' (6.1.3, 6.3 of the present book). It merely means that we would all admit that a person ought to equally treat his own good and another person's good according to their weight, *when he goes beyond his own good and takes the impartial point of view* – in other words, when he takes the viewpoint that he regards his own good as a part of people's good on the whole. Certainly, *if* one takes this impartial perspective, one is required, by the Principle of Benevolence, to equally treat various goods, whether one's own or others'; and if we interpret good as pleasure, one will be led to utilitarianism. However, those who profess egoism can avoid utilitarianism, simply by not taking this impartial viewpoint. In other words, egoists can be consistent in claiming that promoting general happiness is certainly the right thing *if* they take the point of view of the universe, while not taking such a perspective. Sidgwick argues that, when they claim they will not take such an impartial viewpoint because they

emphasize the distinction between themselves and others, we cannot prove their claim to be false.

> When, however, the Egoist puts forward, implicitly or explicitly, the proposition that his happiness or pleasure is Good, not only *for him* **but from the point of view of the Universe**, – as (e.g.) by saying that 'nature designed him to seek his own happiness', – it then becomes relevant to point out to him that *his* happiness cannot be a more important part of Good, **taken universally**, than the equal happiness of any other person. And thus, starting with his own principle, he may be brought to accept Universal happiness or pleasure as that which is absolutely and without qualification Good or Desirable: as an end, therefore, to which the action of a reasonable agent as such ought to be directed.
>
> This, it will be remembered, is the reasoning that I used in chap. xiii. of the preceding book in exhibiting the principle of Rational Benevolence as one of the few Intuitions which stand the test of rigorous criticism.
>
> (*ME* Bk. 4 Ch. 2 Sec. 1 pp. 420–1. *Italics* as in the original text. Emphasis in bold added.)

> In chap. ii. of this Book [i.e., Book VI] we have discussed the rational process (called by a stretch of language 'proof') by which one who holds it reasonable to aim at his own greatest happiness may be determined to take Universal Happiness instead, as his ultimate standard of right conduct. We have seen, however, application of this process requires that the Egoist should affirm, implicitly or explicitly, that his own greatest happiness is not merely the rational ultimate end for himself, but a part of Universal Good: and he may avoid the proof of Utilitarianism by declining to affirm this. It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental, and that consequently 'I' am concerned with the quality of my existence as an individual in a sense, fundamentally important, in which I am not concerned with the quality of the existence of other individuals: and this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual.
>
> (*ME* Bk. 4 Concluding Chapter, pp. 497–8)

Thus, egoism can survive all the arguments which Sidgwick developed as to the theoretical foundations of ethics. We cannot say that

utilitarianism has a more solid theoretical basis than egoism, for both are equally firmly supported by plausible fundamental principles and the claim of hedonism. This being the case, the only viable method of persuading an egoist to pursue people's overall happiness would be to convince him that the pursuit of people's happiness eventually benefits him. Nevertheless, Sidgwick observes that the pursuit of people's happiness does not perfectly harmonize with one's own happiness. The moral acts that utilitarianism dictates and the acts that egoism orders do not always coincide.

As for the relationship between egoism and utilitarianism, Sidgwick admits, on the one hand, that (1) in most societies, to act morally is likely to bring about the agent's own happiness in the long run. This sounds even more plausible if we add that those who perform immoral acts usually run the risk of being punished or sanctioned. Sidgwick also admits that (2) people often feel pleasure by sympathizing with others, and that (3) overly selfish persons tend to have limited interests and therefore are precluded from enjoying various types of pleasures. For all that, he states that the actions which utilitarians dictate may sometimes run against one's self-interest (*ME* concluding chapter, sections 2 and 3). Though it is *generally* disadvantageous for a person to perform an immoral act because of possible punishment, there are cases in which the benefits of acting immorally seem to outweigh the costs because there is little chance of being detected. Some people claim that immoral acts will not make the agent happy because he will be tortured by remorse or by reproach from others, if not punished in public; but there are people who do not feel much remorse, or who do not really care about censure from others (see *ME* Bk.2 Ch. 5). There are other cases in which self-interest and people's happiness come into conflict. A person may confront a situation in which an act to promote people's happiness would force him or his loved ones to make considerable sacrifices. There may also be cases in which a person can best promote people's overall happiness by performing an act which will not bring about his own happiness – for example, by working hard in a solitary place, by working in a stressful environment, or by carrying out an act that would make his loved ones suffer.

Some people, who believe in the perfect concurrence of people's happiness and an individual's happiness, may claim that such harmony is assured by God, namely, that those who do not fulfill their duties will be sanctioned and those who do will be rewarded by God at least in the next world. However, Sidgwick rigidly refuses to accept this as a philosophical claim, maintaining that we cannot find any empirical or

intuitive evidence for the existence of God. From our own experiences on this earth, we certainly see cases in which the pursuit of people's happiness comes into conflict with the pursuit of self-interest. When the dictates of egoism and utilitarianism severely contradict each other, we cannot reasonably decide which dictate we ought to follow. Here arises the problem of the 'dualism of practical reason' which was outlined in Chapter 2 of this book. On this account, at the end of the *Methods of Ethics* Sidgwick had to confess that he could not establish utilitarianism as the *sole* valid ethical theory – though he surely succeeded in showing that utilitarianism has a solid theoretical structure based on rigorous philosophical analyses, undeniable fundamental principles and well-founded claim of hedonism, all of which are supported by our reflective commonsense.

Part II
A Reexamination of Contemporary Utilitarianism

In Part II of this book, we will turn to contemporary moral philosophers – by this I mean twentieth-century philosophers who have reconstructed modern versions of utilitarianism – and reexamine their claims from the viewpoint of Sidgwick's ethics. The following four points will be addressed, as it is commonly believed that these are the main problems Sidgwick left unsolved and that contemporary utilitarians developed new arguments to resolve those problems.

(1) *'Self-evidence' of the intuitive fundamental principles*

Sidgwick presented the Principles of Justice, Rational Self-Love and Rational Benevolence as self-evident principles that serve as the foundation of ethics. According to him, these three principles are apprehended by sophisticated 'philosophical' intuition. However, some people may think it is implausible for us to 'intuitively apprehend' the Principle of Justice, which is that we should not make different ethical judgments about two individuals without any reason, or the Principle of Self-Love, that is, that we ought to equally treat people's goods at different times according to their magnitude. They may also doubt Sidgwick's claim that we can ascertain the self-evidence of such intuitions by confirming that every reflective individual would admit their truth and by observing that such intuitions are also found in our commonsense. R. M. Hare, who was quite critical of the validity of 'intuitions' to which we might adhere erroneously, attempted to develop a contemporary version of utilitarianism without appealing to moral intuitions. To construct his ethical theory, he tried to start with two minimum bases – *logic*, which we all have in common, and *facts*, which we can readily observe. We will examine Hare's derivation of utilitarianism in Chapter 8 of this book.

(2) *Criticism of hedonism*

Some critics might also question the validity of Sidgwick's 'proof' of ethical hedonism. His strategy was to narrow the candidates for the ultimate good down to one by appealing to his own reflective intuition, and to conclude that pleasure is the most plausible candidate for what is intrinsically desirable. However, this was not a decisive argument. This conclusion was only reached by counting on Sidgwick's own intuition and observation.

Of course, Sidgwick made due efforts to confirm the validity of his proof by appealing to other individual's intuition and to people's commonsense. Nevertheless, Sidgwick himself admitted that we cannot reach a complete consensus to approve ethical hedonism. After all, the claim that pleasure is the sole ultimate good stays within the realm of hypothesis. Thus Sidgwick finally appealed to the argument that we cannot find any other theory that can systematically guide our practical decisions. This means that, if we *can* find a theory of the ultimate good which will provide a more systematic guide than hedonism and which will be universally agreed upon, we may well adopt it in place of hedonism. Contemporary utilitarians, such as R. M. Hare, claim that there exists such a theory. According to him, it is not pleasure or happiness but desire- or preference-satisfaction that is to be regarded as the ultimate end of moral actions, or 'the ultimate good' in the area of ethics that deals with what one ought to do. We will examine Hare's preference-satisfaction theory in Chapter 9 of this book.

(3) *Is the sum total maximization principle established?*

Furthermore, there is a question of whether the utilitarian principle of 'maximizing the sum total of people's good' has actually been established. Some critics doubt the possibility of comparing the size of different 'goods' in a way that the concept of 'the sum total of goods' makes full sense. If we accept the Principles of Self-love and Benevolence and agree that we ought to equally treat different individuals' goods according to their magnitude, how do we judge that a person's good and another's good are 'of equal weight'? Sidgwick, who adopts ethical hedonism, claims that the amount of good is to be measured in terms of the greatness of pleasure it produces. He then maintains that the magnitude of pleasure is to be expressed by the strength of desire that one would potentially have at the time of experiencing that pleasure, on the condition that one exclusively considers that feeling. If we admit

his claim, it will follow that *my present* preference can only determine the amount of pleasure that *I* feel *at present*. The amount of my *future* pleasure can be determined only by my preference *at that moment*, and the amount of *someone else's* pleasure can be determined only by *his* preference at the time of his feeling it. Then, *who* on earth can compare the amount of *my present* pleasure with that of *someone else's* pleasure? On what grounds can we say that they are of equal weight?

To avoid this problem, Sidgwick posits a hypothesis that the desirability of pleasure can be known to the individual experiencing it as having a certain definite value, and that to some extent this value can be known to others at different times. We can deny this hypothesis, however. If we deny it, we cannot put equal weight on the various pleasures of different individuals at different times 'according to their size', and thus we cannot form the concept of 'the sum total' of people's pleasures. This kind of problem would arise even if we do not adopt hedonism. To make an ethical judgment in which we must consider the various 'goods' of different individuals from an impartial point of view, we must confront the question of *who* is comparing my present good and someone else's good and *how* we should weigh them. This problem of interpersonal comparison is usually considered to threaten the utilitarian idea of the aggregation of people's good. Chapter 10 of this book will deal with this and other related problems. Some arguments presented by Kenneth J. Arrow and John C. Harsanyi will be mainly considered there.

(4) *Reconciliation between egoism and utilitarianism*

The final difficulty we have to face is 'the dualism of practical reason', namely the problem of the incompatibility between egoism and utilitarianism. According to Sidgwick, a person who simultaneously adheres to egoism and to utilitarianism must fall into a conflict, because these methods of reasoning will sometimes give him mutually exclusive dictates in a particular situation. In such a conflict, he cannot find a well-founded answer as to what he ought to do.

This incompatibility between egoism and utilitarianism is disruptive not only because it causes a conflict in one's own mind but also because it could make ethical discussions futile. If we are about to make a practical decision on a certain social issue, and if the egoistic method and utilitarian method dictate conflicting actions, those who adhere to egoism can choose to endorse selfish actions by not taking a universal viewpoint. Thus, when egoism and utilitarianism give different answers to a particular ethical issue, people's opinions can be divided irreconcilably.

Furthermore, this problem colors the alleged 'self-evidence' of the two intuitive principles that have been offered as the theoretical basis of egoism and utilitarianism. The possibility that egoism and utilitarianism may give contradictory guides for actions indicates that the Principle of Rational Self-Love applied to one individual and the Principle of Rational Benevolence applied to people in general can give conflicting dictates about what ought to be done in a particular situation. This does not mean that the Principles of Self-Love and Benevolence directly clash – they do not logically contradict each other, and, in fact, utilitarianism is composed of both of these principles. Still, these two intuitive principles certainly cause practical conflict when applied to a particular situation. Therefore one may well doubt whether these two principles fulfill one of the necessary conditions for 'self-evident and significant propositions' that Sidgwick had presented, that is, the condition of consistency (see 5.1.3 of the present book). If not, then one or the other or both of these principles of Self-Love and Benevolence would turn out to be false as self-evident and significant intuitive principles. This would cause a fatal flaw in Sidgwick's whole ethical framework.[1]

Even if we deny this possible clash between these two intuitive principles, to leave practical conflicts caused by the egoistic and utilitarian methods would be against the fundamental postulate of ethics, which states that 'so far as two methods conflict, one or other of them must be modified or rejected' (see 2.1.2 of the present book). By admitting the conflict between the two methods, we are compelled to admit that our reasoning in deciding what ought to be done cannot be completely consistent and free of conflict, and that ethics cannot be fully systematized.

In fact, when he had to leave this problem of practical dualism unsolved, Sidgwick felt he had finally failed to completely systematize ethics, so he concluded the first edition of *The Methods of Ethics* with the term 'failure' (*ME*1 p. 473). At the end of the seventh edition he still made an equivocal remark about the overall consistency of ethics, suggesting that to abandon the hypothesis that ethics can become systematic and consistent would be to open the door to a 'universal skepticism' of every phenomenon in the universe (*ME*7 p. 509). If we regard this dualism as a failure of ethics itself, or as causing skepticism about ethics, we may well suspect the theoretical validity of utilitarianism, which we have been trying to establish. We must somehow solve this problem of dualism and seek a way to prove the supremacy of utilitarian ethical

theory, which Sidgwick personally upheld. Richard B. Brandt also dealt with this problem. He tried to show that what we would all support as 'a social moral system', that is, a regulatory system that should be widely accepted to control the acts of each member of a society, is utilitarianism. Chapter 11 of this book will examine Brandt's attempt.

This page intentionally left blank

8

An Approach not Appealing to Moral Intuition

Sidgwick used his philosophical intuition to reach the three fundamental principles, from which he developed his argument about the foundation of utilitarianism. The reason why Sidgwick believed we could rely on these three 'intuitive' principles, while dismissing perceptional and dogmatic intuitionism, was because these three principles use clear and definite terms, because their validity can be repeatedly confirmed by reflection, and because these are presumably accepted by most or all people regardless of the differences in the ethical views they usually hold. However, we may question whether his 'philosophical intuition' is really common to us all, and on what grounds we can say so. In replying to this problem, the moral philosopher Richard Mervyn Hare (1919–2002) attempted to develop an ethical theory by appealing only to *logic* and *facts*, which we would surely accept as our common basis, and which we can ascertain as being universally valid by observing how we actually behave and how we actually use our language. His argument is also important in that it led him to advocate a version of utilitarianism, taking quite a different route from Sidgwick's. In this chapter, we will examine Hare's argument for utilitarianism and compare it with Sidgwick's analyses.

We should remember here, however, that Sidgwick and Hare hold different versions of utilitarianism – Hare supports what is called preference-utilitarianism instead of hedonistic utilitarianism. According to preference-utilitarianism, the morally right action is the one that aims to maximize not the happiness but the *preference-satisfaction* of all parties concerned. We will discuss this difference in the next chapter. In the present chapter, we will focus on the question of how Hare's argument produces the basic utilitarian claim of 'equal consideration for, and the maximization of the sum total of, all people's goods'. In doing so, we will examine whether Hare has successfully avoided the difficulty that

Sidgwick had to face when he tried to base utilitarianism on the three 'intuitive' principles. Contrary to common perception, I will argue that Sidgwick was more accurate than Hare in analyzing the logic of our ethical judgments and the foundation of utilitarianism.

Hare usually calls his study 'moral philosophy', which deals with moral thinking about important practical issues (*MT* Preface and others), or with rational thinking about moral questions (*MT* 12.4). A person has a moral question, and is doing moral thinking, when he asks 'what (morally) ought I now to do?' (*MT* 12.4, p. 214). We have previously distinguished 'ethics' in a wide sense and 'morality' in a narrower sense (1.1 and 3.1 of the present book), and therefore might be a little confused by Hare's use of the parenthetic term '(morally)'. We should note, however, that Hare defines the scope of 'moral questions' simply as (1) the questions that involve value judgments which are expressed by such terms as 'ought' and 'right', and as (2) those which are especially concerned with certain value judgments that always override other value judgments – Hare regards this characteristic of 'overridingness' as what demarcates *moral* judgments from other types of evaluative judgments (see *MT* 3.5 and 3.6). Hare believes it would be safe for us to assume that characteristics (1) and (2) are the basic features of the problems, decisions and conduct which are called 'moral', and he says nothing more to define the notion of morality. Hare places such minimum restraints on the scope of morality because to further qualify what is moral would be to arbitrarily narrow the scope of the problems we need to deal with. However the problems are defined, Hare aims to deal with those in which we ask what we ought to do, and which have overriding importance. Thus, we can assume that Hare's moral philosophy deals with issues as broad as Sidgwick's ethics. Like Sidgwick, Hare deals with every kind of normative judgment about an individual's actions, that is, the judgments about what an individual ought to do (in other words, which action he ought to choose to perform by voluntary effort). We should keep in mind, however, that Hare's chief concern is with moral issues that arise in a situation in which the interests of others are affected by one's decision (*MT* 3.5 p. 54). Furthermore, we will later notice that Hare sometimes uses the term 'moral' with some additional connotations, which will be explained later.

8.1 Hare's stance

What is remarkable for Hare's moral philosophy is that he devised a method of attaining a universally acceptable moral judgment without appealing to the moral intuitions of particular individuals.

What Hare calls 'moral intuition' is the intuition that apprehends substantial moral rules or principles as self-evident. *Moral rules or principles* are the rules or principles that guide us to determine actions to be done. When a moral rule is a *substantial* one, it more or less concretely specifies the *substance* of the action that ought to be performed. Moral intuition convinces a person that certain substantial moral rules or principles are obviously valid without the need of proof. Such moral intuitions include what Sidgwick called dogmatic intuition, namely, the intuition that apprehends relatively simple and brief rules – for example, 'do not tell a lie', 'do not steal', and so on – as self-evident. However, they also include other types of intuitions that presuppose certain normative judgments as self-evident, such as the ones John Rawls, a target of Hare's criticism, appeals to at every crucial point of his arguments. Hare does not explicitly state to which moral intuitions Rawls erroneously adheres, but presumably they include Rawls's intuitive attitude that never doubts the unconditional rightness of his difference principle, which presupposes that those most disadvantaged should always be given priority in the distribution of social and economic goods[1] (*MT* 4.4, p. 75).

While admitting that moral intuitions play an important role in our daily life, Hare argues that we should never appeal to them when we are dealing with very serious moral issues.

Certainly, there are several merits in having sound moral intuitions firmly inculcated in our mind. In so far as it is a sound one, a moral intuition enables us to instantly perform proper actions in emergencies where we have no time for deliberation. It also enables us to resolutely decide what we ought to do without succumbing to temporary temptations or predicaments. However, we cannot always rely on our moral intuitions to decide what we really ought to do. Most of such intuitions are what we have learned as the result of past education or experience, but our past edification may not always have been correct. We may confront a novel situation and become bewildered not knowing which intuition we need to apply. Sometimes our moral intuitions may conflict with each other. In all such cases, we are propelled to reconsider the validity of our own moral intuitions. However, it will become a vicious circle if we try to solve these difficulties by appealing to our own moral intuition. When the very intuition is in question, it would be futile to answer that 'my intuition would be correct' without presenting any further grounds beyond one's original conviction in its validity. Especially when there is a moral conflict among people, the argument that appeals to moral intuition would only bring about a reconciliation among those

who already share that intuition – for moral intuitions have 'no probative force, and the two sides in the most important moral arguments will have different intuitions' (*MT* 1.3, p. 12).

These suggestions overlap Sidgwick's argument on dogmatic intuitionism. Sidgwick also claimed that, though our dogmatic intuitions are by and large sound, they contain ambiguities and may often raise conflicts and doubts, and that a truly systematic ethical theory must therefore have a more profound basis than dogmatic intuitions. However, Hare takes a more rigorous stance than Sidgwick, attempting to exclude *any* appeal to moral intuitions in developing his ethical theory.[2] Even if a person proposes a more refined philosophical principle than existing moral rules, we cannot say that it is truly valid as long as his grounds for supporting that principle is merely that its truth is intuitively self-evident to him. Even when that intuition is widely shared by people, it might be that the whole group is deceived by erroneous information or fallacious beliefs. In any case, when people's opinions differ, or when we happen to have doubts about our own moral intuitions, we cannot solve a problem by claiming that we have been convinced that this moral intuition is correct and valid.

To sum up, when we have a serious moral issue for which we cannot bring about a substantial solution by appealing to moral intuitions, we cannot rely on 'a moral truth' the grounds of which are unknown, or on 'a fundamental moral principle which my Reason seems to have apprehended' and on which there is no guarantee that people have reached a consensus. Then, how can we rationally decide what we ought to do without appealing to *any* moral intuition?

With this question in mind, Hare directed his attention to *logic* and *observable facts*, as the minimum materials that we need to make moral judgments.

First, whatever judgment we may make, it is necessary to follow the logic that governs it. The kinds of reasoning about how to judge what the facts *are*, what one *shall* do, and what one *ought* to do, are all very different. Such differences partly depend on the logic particular to each judgment. An individual can make any type of judgment of his own will, but if he is going to make a moral judgment he cannot deviate from the logic of moral judgments – if he deviates, then his judgment will not be regarded as a moral one.

When we make a moral judgment, we make a judgment which can be expressed by such phrases as 'I *ought* to do such and such', ' such and such is the *right* action', etc. Though the content of moral judgments varies, one thing is certain: every moral judgment commonly

uses such terms as 'ought' and 'right'. Thus, if we are going to make a moral judgment, we are required to first recognize the meaning and use of these terms, and understand the logical properties of ought- or right-judgments. When we understand those properties, we can tell when we deviate from the logic of moral judgments, and can grasp the logical requirements that we must fulfill in order to make a moral judgment.

Hare states that the meaning and use of a term can be known by examining our 'intuitive' recognition or behavior about language. He calls this intuition 'linguistic intuition'. It was previously stated that we should never rely on our moral intuitions, but that was because it is the differences in our moral intuitions that cause moral conflicts or disputes. By contrast, linguistic intuition remains the common conceptual basis among people even when we set aside all moral intuitions that can cause antagonism among us. Since language exists mainly for communication among people, there usually is a broad consensus on the general meaning and use of the words in it (*MT* 1.3, p. 11 f.). Therefore, Hare believes, we may regard our linguistic intuition as the basis of logic. We could use artificial language instead of the existing one if that helps us solve the moral problem before us. However, Hare thinks that it would be more feasible, and helpful, for us simply to clarify the terms and concepts that we currently use rather than to adopt brand new language.

We cannot make a moral judgment just by elucidating its formal elements, such as logic, however. A moral judgment also has empirical content, which relates to the real world. If we are to make a practical judgment that applies to an actual situation, we should not ignore the facts about that situation and the people in it (*MT* 1.2, 3.2, 5.1, 9.9). This does not mean that we must apprehend some extraordinarily profound truths or supernatural facts: it suffices if we recognize ordinary facts that nobody would deny. Hare never imports so-called moral facts into his arguments, for it would be to introduce his moral intuitions. A moral judgment must have certain substance, but this substance is not predetermined. We determine it *after* we have considered various common facts.

Now, do we need a third element, besides *logic* that determines the form of moral judgments and *empirical facts* that constitute the substance of moral judgments, in order to make a moral judgment? Hare admits that a moral judgment cannot be made without a *prescriptive* element (explained later), which indicates the voluntary will of the individual making a judgment; nevertheless, he attempts to show that 'if we assumed a perfect command of logic and of the facts, they would

constrain so severely the moral evaluations that we can make, that in practice we would be bound all to agree to the same ones' (*MT* 1.2). According to Hare, by carefully following logic and fully recognizing the facts we can not only make a coherent moral judgment without appealing to moral intuitions but also reach a *unanimous* judgment. This would mean that we could come to a settlement on a moral issue, despite the differences in moral intuitions we usually have. Thus Hare attempts to construct a moral theory using only two elements, logic and facts (plus the will of the person making a judgment). Hare tries to completely avoid other elements than these two. He never introduces 'moral truth that our reason intuitively apprehends' into his argument; for such a 'truth' can perhaps be one individual's disguised moral intuition. For such reasons, Hare gives to the term 'rational' only the minimum meaning of 'having fully recognized the facts and followed the logic in thought directed to the answering of questions' (see *MT* 12.4).

8.2 Hare's argument for utilitarianism

However, this is not the end of Hare's argument. Surprisingly enough, he further maintains that a moral theory supported by his method, which uses only logic and facts, must be a version of utilitarianism. According to him, if we admit that we are going to make a moral judgment *without appealing to our moral intuitions*, if we observe the logic of moral judgments, and if we precisely recognize relevant facts, we will attain a utilitarian moral judgment, in which we choose an action that will bring about the greatest preference-satisfaction for all parties concerned.

Then, what is the logic of moral judgments, and what kinds of facts should we recognize when we make a moral judgment? How can utilitarianism be established from them? In the following, I will examine several points in Hare's argument which we may find bewildering, compare them with Sidgwick's arguments on the three fundamental principles, and articulate the weakness of Hare's derivation of utilitarianism.

8.2.1 Logic of moral judgment

According to Hare's analysis, we can suggest two logical properties of moral judgments, which are *prescriptivity* and *universalizability*.

Prescriptivity is the property of implying one or more imperatives, or prescriptions (*MT* 1.6). A prescription is a kind of statement used to guide our action. It does not have enough power to compel people to perform a prescribed action even when they do not agree to the prescription in

question. However, it would be contradictory if a person sincerely agrees to the prescription 'Do X in a certain situation' and yet does not perform X in that situation.

What Hare means by talking about 'the prescriptivity of moral judgments' is that moral judgments, in which the term 'ought' is used, also contain a certain kind of prescription. Unlike a descriptive judgment that describes and reports a certain fact, an ought-judgment functions as a prescription to bring about a certain fact. When I judge that person P *ought* to do action X in situation S, I am recommending action X to him and prescribing him to do that act in situation S. Of course, this 'person P' can be the very person making this utterance. When I judge that *I* ought to do action X in situation S, I am prescribing to myself that the action be done in that situation. In any case, to agree to an ought-statement is to agree to the prescription it involves; and as long as it is a sincere one, a person who has agreed to this statement must be ready to perform action X in a situation where this prescription applies to himself.

That ought-judgments have prescriptivity can be ascertained by the following linguistic intuitions and behaviors of our own. (More precisely, we can explain our intuitive thoughts and responses as to our linguistic behaviors and our use of words by postulating that ought-judgments have prescriptivity. Though this logical property is only a hypothetical one, Hare thinks that as far as it corresponds to facts this hypothesis can be retained without being disproved.) First, we use ought-judgments when we give advice or instruction to someone, or when we are deciding what to do. If ought-judgments did not have prescriptivity – for instance, if ought-judgments were judgments that merely described one's own internal emotions – we would not be able to understand why we use ought-judgments in such circumstances. Second, if a person who has agreed to the statement 'I ought to do this action now' does not actually perform that action when it is physically and psychologically possible for him to do so, we will think that *either* his judgment was a insincere one, *or* he does not know how to use, or simply misuses, the term 'ought'. Hare admits that judgments containing the term 'ought' do not always entail imperatives or prescriptions in this way; for example, the ought-judgments that merely refer to existing moral rules in a society (for example, when a serial killer grins and says 'yes, we *ought* to respect the life of others'), or those which only express one's moral sentiments often lack prescriptivity (*MT* 3.7). Admitting that there are such exceptions, Hare insists that there are certainly cases in which ought-judgments have prescriptivity. In particular, when we

face serious moral problems and consider what we really ought to do, the ought-judgment we make must have prescriptivity. Our linguistic intuition tells us it is true; nobody would deny that a person is showing a logical inconsistency when he makes a judgment about what he ought to do in such a serious situation and yet does not actually act according to his own judgment.

Now that we have ascertained that ought-judgments have prescriptivity, we need next to make the point that the prescription which an ought-judgment contains is an expression of the *preference* of the person who is making this judgment. A person voluntarily makes a prescriptive ought-judgment, and this prescribes what he ought to do, only when he assents to this judgment and the prescription it implies. Whether he assents to this prescription depends on the preference of the assenting individual, and the fact that he prescribes it indicates that he prefers it. Thus, the prescription implied in an ought-judgment is based on the preference of the person making that judgment. Prescriptivity of an ought-judgment is generated when an individual prefers a certain act to other alternatives and recommends its accomplishment.

However, we have to clarify one more point for the sake of the subsequent arguments, which is that an ought-judgment and the prescription it implies are the expression of the *final* preference that an individual would have at the time of making that judgment. This preference may not be what he originally held before he started to consider which judgment to make, but can be the one he reaches after a certain period of deliberation. A person who makes a judgment may experience an internal conflict among various preferences of his own before he reaches the final decision. A person can have multiple preferences at the same time – he may want to save money on the one hand, but may be inclined to spend extravagantly for his vacation on the other. If he is to make a judgment about what he ought to do in a certain situation while having multiple preferences about it, he would weigh all the preferences that he has, in order to consider whether he can really comply with the prescription that each preference implies; then he would finally decide on a certain ought-judgment, resolving to assent to the prescription it entails. The prescription implied in that judgment is the expression of the *final* preference of a person making that judgment, and it does not directly reflect various preferences he had and considered before he reached his final decision. This point will become very important later.

Thus we can at least say that serious ought-judgments are prescriptive, and their prescriptivity represents the final preference of the persons making those judgments. From this logical property of prescriptivity one

requirement we must meet in order to make a rational ought-judgment (that is, one that conforms with logic and facts) becomes clear: If you are to make a certain ought-judgment, which can be expressed as 'person P *ought* to do act A in situation S', you must admit that you are *prescribing* him to do A in situation S. If this judgment is expected to apply to yourself, you must be able to *prescribe yourself* to do A in situation S, and you must *be ready to do* A in that situation. You must *prefer* to make that ought-judgment, resolving to accept all the prescriptions it entails. I will call this ***the requirement of prescriptivity***.

The second logical property of moral judgments is, according to Hare, ***universalizability***. It is the property that a particular ought-judgment invokes or implies a judgment which does not contain any reference to a particular individual. Hare calls this latter judgment a *universal* judgment,[3] and hence the former a *universalizable* judgment (*LM* 10.3, p. 156; *FR* 3.8, 9.2). Unlike simple imperatives or desires, to make an ought-judgment involves a commitment beyond making that particular judgment. This difference between ought-judgments and simple imperatives or desires is universalizability. Even if a person makes an ought-judgment about a single situation in which a particular individual is involved, by making this particular judgment he is explicitly or implicitly showing his commitment to a universal judgment which can be applied to all similar individuals in similar situations.

That a judgment does not contain any reference to a particular individual means that it does not refer to any particular individual person/object, nor to any particular point in time. An ought-judgment must, even if it is made primarily for a particular person at a particular point of time, presuppose a universal judgment that does not especially single out that person or that time. Hare points out that a particular ought-judgment such as 'I ought not to kill individual A in this situation' always implies a judgment which can be expressed as 'any individual similar to me ought not to kill any individual similar to A in a situation similar to this'. Though the latter judgment apparently mentions the particulars, 'I', 'A' and 'this', Hare says that the latter judgment does not single out any of them (for they are mentioned just as one among many that belong to the same group), and that therefore we can regard the former particular judgment as universalizable.

That 'an ought-judgment invokes or implies a universal judgment' entails that, when a person makes an ought-judgment as to situation S, he is committed to make the same judgment about *all* similar situations whose universal features he admits to be equal to the ones in situation S (*MT* 6.4, p. 115). '*All* similar situations whose universal features are

equal' include, according to Hare, all hypothetical situations which are similar to the actual one in question *except* that the roles played in them by particular individuals are different. Thus they also include an imaginary situation which is similar to the actual one but differs only in that the positions of myself and someone else are hypothetically reversed. Hare claims this because merely to exchange a particular individual for another causes no change in the universal features of the situation. If this is correct, to make an ought-judgment in a particular situation is to show one's commitment to make the same kind of judgments about all similar situations, including those in which the positions of the individuals concerned are hypothetically exchanged, and hence including the one in which *I* am hypothetically put in someone else's place. Thus, when I make a judgment, for example, that 'Doctor A ought to give life-prolonging treatment to patient B who has the will to live longer' in one situation, I am simultaneously showing my commitment to make the judgment that 'even if the patient were A and the doctor were B, the doctor ought to give life-prolonging treatment to the patient who has the will to live longer, as long as there is no further difference in the described situation'. This imaginary reversal of positions includes, of course, the exchange of my position with that of someone else. Therefore, if I claim that *I* ought to do something to someone, I am simultaneously bound to admit that *if I were in his position* the same thing ought to be done to me.

According to Hare, that ought-judgments have universalizability can also be ascertained by our linguistic intuitions and behavior. (More precisely, we can explain human behavior by postulating that our ought-judgments have universalizabilty.) We think that a person is showing logical inconsistency if he makes different ought-judgments about cases which he himself admits to be identical in their universal features. To use Hare's own words, if a person states that 'You ought, but I can conceive of another situation, identical in all its properties to this one, except that the corresponding person ought not' or that 'Jack did just the same as Jim, in just the same circumstances, and they are just the same sort of people, but Jack did what he ought and Jim did what he ought not', we will be perplexed, just as when we hear the statement that 'The two figures are exactly the same shape, but one is triangular and the other not' (*MT* 1.2, p. 10; 4.7, p. 81; 6.4, p. 116). We will simply be unable to understand him, or we will believe that he is either being dishonest or misusing the term 'ought'.

From this logical property of universalizability, we can obtain another requirement that we must fulfill in order to make ought-judgments

meaningfully: When a person is to make a certain ought-judgment about a certain situation, he must be prepared to make the same judgments about similar situations having the same universal properties. I will call this ***the requirement of universalizability***. The situations considered here include all hypothetical ones in which a person making a judgment is put in the same position as the one that someone else occupies in the actual circumstance. Therefore, when I make an ought-judgment in a certain situation, I have to consider if I can make the same ought-judgments about all similar situations, including the hypothetical ones in which I imagine I am put in the position of someone else who is involved in this situation.

So far I have explained Hare's analysis of moral judgments, ascertaining their two logical properties and suggesting that in making an ought-judgment we must satisfy both the requirements of prescriptivity and universalizability.

At this point let us temporarily depart from Hare's own argument, and consider several points that may draw our attention. Needless to say, Hare's prescriptivity and universalizability correspond to what Sidgwick suggested in his analyses of 'ought' and 'right', and Hare's requirement of universalizability is essentially the same as Sidgwick's Principle of Justice. According to the Principle of Justice, we have to make the same ought-judgments for two similar situations which involve two individuals unless there is a difference in the nature or circumstances of those individuals apart from being different people. We also have to make the same judgment about the two similar situations in which two individuals reverse their positions, other things being equal. This is essentially the same requirement as Hare presents, since this principle is proposed as a guide we must follow when we make judgments that contain such terms as right or ought. However, there are several, seemingly subtle, differences. First, Hare clearly emphasizes that the requirement of universalizability is only meant as a condition for following the logic of ought-judgments, and that it contains no further connotation as to our moral obligation. According to Hare, we do not need to claim that this requirement is 'an intuitive truth apprehended by our reason'. We actually make ought-judgments in a moral context, regarding them as having universalizability; and therefore a person who intends to make such a universalizable ought-judgment should not deviate from the logic of universalizability. According to Hare, this is all that universalizability requires.

Second, Sidgwick's Principle of Justice concerns the *logical* treatment of different individuals, in the sense that it requires us to equally treat

different individuals, by applying the same ought-judgment to them all. For Sidgwick, the requirements regarding the treatment of *quantity or strength of preferences of different individuals at different times* are concerned with still other principles, that is, the Principles of Self-Love and Benevolence, which require a certain consideration of *the amount of good* (see Ch. 6 of this book). We should note that, in Hare's theory, any principle about such a quantitative evaluation is not introduced throughout his meta-ethical analysis of ought-judgments. This point is very important. Hare's requirement of universalizability itself is about the logical treatment of different situations and individuals, which only requires us to make the same ought-judgments about all the similar situations and similar individuals, regardless of the mere differences in time and in the roles played by the particular individuals. It does not state anything about the treatment of *quantitative* things, such as the weighing of goods or preferences. What Hare claims is only that we must make the same ought-judgment about similar situations in which my position and that of someone else are exchanged but other than that we can find no other significant differences in the nature and the circumstances of the individuals involved. If this understanding is correct, I will be fulfilling both Sidgwick's Principle of Justice and Hare's requirement of universalizability as long as I am prepared to make the same ought-judgments about all those situations, *whatever weight I put on them*.

We should note, however, that Hare further claims that ought-judgments also have prescriptivity, and that this combination of presctiptivity and universalizability of ought-judgments, plus the precise recognition of facts, lead us to put equal weight on people's preferences. According to him, an ought-judgment is universally prescriptive in the sense that it applies to all similar situations, including ones in which people, including oneself, exchange positions while preserving the situation's universal features, and that it prescribes the same action about all those cases. Thus Hare claims that, in making an ought-judgment, a person first has to recognize that the prescription it involves would be applied to all similar cases, including one in which he is put in the position that someone else currently occupies, and also has to form his final preference to make that judgment (see, for example, *MT* 5.1 p. 89). Despite this argument, I would still emphasize the following. Even if I temporarily – that is, before I make my final ought-judgment – feel reluctant to accept the prescription that a certain ought-judgment entails, as long as I finally resolve to make a prescriptive judgment and to undertake the prescription it involves, I am satisfying both of the two logical requirements for ought-judgments. If I am aware that

by doing so I am showing my commitment to accept similar prescriptions for all similar situations in which I may be placed in another's position, I am satisfying the requirements of both prescriptivity and universalizability.

8.2.2 Relevant facts

Let us get back to Hare's own arguments. After clarifying two logical properties of ought-judgments, he suggests that in order to make a rational ought-judgment we need not only to be logical but also to recognize relevant facts (*MT* 5.1). First, we usually make ought-judgments based on certain facts. For example, when people claim that euthanasia ought to be permitted because patients are often in severe physical pain, their judgment is based on the fact that some patients are in so much pain that they sincerely wish to die. Therefore, we need to ascertain if the alleged facts are true by precisely recognizing the situation in question. Also important is the recognition of the expected consequences of making that judgment. Whatever kind of prescription we make, it is irrational to prescribe without recognizing exactly what one is going to prescribe. In order to offer a helpful and practical guide for action, we need to know what effects or consequences this guide would actually bring to the situation in question. To precisely predict such effects and consequences, we need to have sufficient knowledge of the actual situation in question. In addition, we would also admit that our choice of moral action or principle depends on alternatives. We need to know what alternatives we have as to possible actions and their consequences. Such a consideration of alternatives becomes more exact as we obtain more detailed knowledge of the facts about the actual situation.

However, there is a more important class of facts that we need to know in making moral judgments. Hare claims that the 'obvious candidates' for such notable facts are those which tell us the probable effects of possible actions on the preference-satisfactions of people (*MT* 5.2). Thus we need to know what preferences people actually have or are likely to have.

Probably most of us will understand from the above explanation that we need to know present and expected facts about the consequences of actions. However, why do we need to recognize the facts about *people's* preferences? Certainly, the prescriptivity of ought-judgments would suggest that an ought-judgment that a person makes must be the one *he* can prescribe and prefer. Therefore, *his* preference is obviously relevant to his ought-judgment; but why should he recognize the facts about other people's preferences?

Here we should note Hare's stance on ethics, or moral philosophy. When he analyzes the logic of moral judgments, Hare stipulates only the logical properties, which everyone who uses the same language would admit. So he seems to be concerned with all kinds of 'moral' issues in which one may want to make a universally prescriptive (and overriding) ought-judgment. However, when discussing moral philosophy Hare actually confines his attention to serious moral issues *in which the interests of others are affected*. He clearly states that 'moral judgments, though they are not confined to situations where the interests of others are affected, have their predominant use in such situations. For cases where the interests of others are not affected, I make no claim to provide canons of moral reasoning' (*MT* 3.5, p. 54).

Some may claim that morality is concerned not with interests but with ideals; a 'problem' arises, however, not when one person's ideal is simply different from someone else's ideal, but when that difference causes conflicts that would damage the second person's interest. This being so, a situation in which a moral issue arises already contains the fact that the interests of others are or will be affected. What a person regards as his interest would be either what he now desires or what he would desire at some future point in time, as well as the means to satisfy such desires. If moral issues, especially serious ones, are concerned with people's interests, those problematic situations already involve, *as a matter of fact*, the desires, likings and preferences of the concerned parties. If a moral issue already involves other people's preferences, which actually exist and can be affected, it would be irrational to make a moral judgment about that issue while ignoring the actual conditions of those preferences and the effects of our decision on them.

Furthermore, the requirement for knowing the facts about others' preferences is intensified because of the universalizability of ought-judgments. If I am to make an ought-judgment about a certain moral issue, I must be able to assent to the universal judgment that does not single out any particular individual. This means that I must be able to make the same judgment in a hypothetical situation whose universal features are identical to the actual situation in question but in which *I* would be placed in a position which someone else currently occupies. Therefore, in order for me to ascertain whether I can really make the same ought-judgment while putting myself in a position which is actually occupied by another, I need to know the facts about the other's actual position. Hare thinks that such facts must include facts about *what preferences that person actually has*. Assuming that he has his own state of mind and his own preferences that are different from my own,

I have to precisely recognize those facts. According to Hare, otherwise I cannot say that I could fully imagine what it is to put myself in the same position as another actually occupies.[4]

There are also other facts that we need to consider. Even within one and the same person, preferences change over time. If one's preferences have changed, the effects of a moral judgment on them also change. The possibility of preference change also affects our feasible alternatives of action. Therefore, if I am to fully recognize all the facts relevant to the ought-judgment I am about to make, I also have to consider the possibility that people's preferences may change in the future, as well as the consequences of such preference change. Hare thinks that we should include our own as well as others' *future* preferences into the 'facts' that we take into consideration. If we could have perfect knowledge of our past, present and future, we would be able to make a perfectly universal judgment in the sense that it does not refer to any particular time or to particular individuals. Though it is practically impossible for us humans to have such perfect knowledge, we can still make a best guess by carefully observing the facts about current situations.

8.2.3 Cognition and replication of preferences

Thus, if Hare is correct, in making an ought-judgment about a situation in which the interests of others are involved, I have to imagine a case in which I am placed in the same position as another actually occupies, and hence I have to know the facts about the preferences that he or she has now, or will have in the future.

At this point Hare further claims that, in order for me to precisely imagine what it would be like for me to be in another's actual position, I must put aside my present knowledge and position. Then I must imagine myself in exactly the same position as he occupies, *having the same feelings and experiences as he actually has*. To put it bluntly, I have to *identify* with him and his own preferences. Hare contends that unless I do so I cannot say I have fully imagined myself in exactly the same position as he occupies. According to Hare, 'another's actual position' must mean his situation including his state of mind, and especially, his preferences. Thus if I am to fully imagine what it is like to be in another's position, I have to put myself in another's shoes with *his* preferences and without the preferences or knowledge I originally had.

> I emphasize that the imagined situation must be one in which I have *his* preferences. If, by some quirk of nature, I were a person who

> knew that he did not feel pain in that situation, or if I knew that I was going to become such a person by being anaesthetized, then I might indeed sincerely say that I did not mind being subjected to the experience (ignoring for the sake of argument its consequences). But this would be irrelevant; and so would it be if I knew that I would feel pain, but for some reason would not mind it. For I am to imagine myself in his situation with *his* preferences.
>
> (*MT* 5.3, p. 94)

Hare regards this state of imagining oneself in another's position as what makes one *absorbed* by another's position, while putting aside one's own present knowledge, preferences and mental states that the other does not actually possess. I will term this claim that 'one may not bring along one's own present knowledge, preferences and states of mind when one puts oneself in another's shoes' ***the Principle of Absorption***. This is not a requirement for following the *logic* of ought-judgments. Rather, it is a requirement for *the precise recognition of facts* that are relevant to the moral judgment one is going to make. When I imagine what it is like to be in another's exact position, I have to be absorbed into his actual situation with his preferences.

With this point in mind, Hare makes another, even more important claim. He insists that, in order for us to know what it is like to be in another's position with *his* preferences, we must *now* represent to ourselves, or replicate in our minds, the preferences that he actually has or will actually have. That is, we must *now* possess the preferences whose quality and intensity are exactly the same as his present or future preferences. Hare claims that otherwise 'I cannot really be knowing, or even believing, that being in his situation with his preferences will be like *that*' (*MT* 5.3, p. 95). When we reflect on our own preferences, we would admit the following: when I have a strong or weak preference for something, I *now feel* that preference with the same intensity, and have the *motive* to satisfy that preference (*MT* 5.2). This is because preference is always prescriptive. Likewise, when I know someone else's preference for something, I must now feel that preference with the same intensity, and acquire a prescription to satisfy that preference – so Hare claims. The preference I newly acquire is not the preference I used to have but the exact copy of *his* preference, namely the preference that has equal quality and intensity as his. As far as it is *a fact* that he has that preference, if I fully recognize it I must have replicated that preference to the same degree. Whether his preference is a good or an evil one is irrelevant to the present argument.

Hare explains this point as a relationship between the following two propositions.

> (1) I now prefer with strength S that if I were in that situation x should happen rather than not;
> (2) If I were in that situation, I would prefer with strength S that x should happen rather than not.
> [. . .] What I am claiming is not that these propositions are identical, but that I cannot know that (2), and what that would be like, without (1) being true, and that this is a conceptual truth, in the sense of 'know' that moral thinking demands.
>
> (*MT* 5.3, pp. 95–6)

Thus Hare asserts that 'I cannot know the extent and quality of others' sufferings and, in general, motivations and preferences without having equal motivations with regard to what should happen to me, were I in their places, with their motivations and preferences' (*MT* 5.4, p. 99). This is what Allan Gibbard named 'the Conditional Reflection Principle' (Seanor and Fotion 1988, p. 58). The same argument can also be applied to one's future preferences. If we fully recognize our future preferences, we should *now* acquire preferences whose quality and intensity are equal to the future ones.

This part of Hare's argument might seem somewhat difficult to understand. Of course, Hare is not claiming that we can actually have perfect knowledge of another's preferences or of our own future preferences. Hare insists, however, that it would not be entirely impossible for one to represent to oneself one's past or future states of mind, or those of others including their preferences. We should remember here that, for most preferences, which can be expressed as 'one prefers A to B', either A or B, or both, must be an experience or a thing which does not yet exist. We usually control our lives by predicting, or representing to ourselves as precisely as possible, a state of affairs which we have not yet experienced. We humans are not endowed with such perfect sensitivity and sympathy that we can fully picture the preferences of others. So we may erroneously imagine preferences that they do not actually have. However, we need to approximate a precise recognition of another's position if we are to make a rational ought-judgment that is fully logical and truly based on facts (see *MT* 7.4).

Nevertheless, it may still seem difficult for us to accept this Conditional Reflection Principle. Hare claims that this is 'a conceptual truth, in the sense of 'know' that moral thinking demands' (*MT* 5.3 p. 96); but is his contention true? When we say we 'know' what it is like to suffer as this

or that person, and yet do not have an equally strong preference that the same suffering should not happen to ourselves, do we regard it as a misuse of the word 'know'? There seems to be no logical inconsistency if I answer in the negative.

As another possible explanation of this 'Conditional Reflection Principle', Hare attempts to explain that the term 'I' might also have prescriptivity (*MT* 5.4). According to this explanation, when I call some person 'I', I am undertaking a commitment to satisfy that person's preferences. I do not necessarily aim to satisfy *his* preferences, but by calling his position 'mine' while I consider making an ought-judgment, I am inclined to satisfy the preferences that *I* would have if I were in that position. Hare conceives that this would be equivalent to having *my own* preferences in that position.

It is still unclear, however, why I have to *entirely* put aside my preferences for the purpose of knowing someone else's preferences while imagining myself in his position, and why I have to make such an almost impossible effort to acquire a precise copy of someone else's preferences. This part of Hare's argument seems too much to ask of us. In fact, the reason why Hare adopts this method of replicating preferences is because he needs to avoid the difficulty of the interpersonal comparison of preferences, which bothered Sidgwick (see 7.2.1 of this book). As will be clarified in the next section, once one successfully creates in one's own mind the copies of other people's various preferences, the interpersonal comparison of preferences can be converted to an *intra*personal one. The usefulness of this method will be discussed in 10.2 of this book.[5]

Perhaps I was a little hasty in suggesting my own evaluation of Hare's argument of the conditional reflection. What we should remember at this point is that Hare's 'conditional reflection' is what we are required to do as the necessary condition for *the precise recognition of relevant facts*, that is, recognition of the actual positions of others. We need to do more than just recognizing those positions in order to reach a final ought-judgment. Hare's argument so far has only led us to now acquire our own preferences equal to those of someone else (or to someone's future preferences), as the preferences regarding a hypothetical situation in which we are put in his present or future position. It does not follow that such acquired preferences, that is, the copies of others' present or future preferences, immediately determine our final moral judgment.

8.2.4 Utilitarian moral judgment

At any rate, according to Hare's argument, if I am going to make a rational ought-judgment, I am supposed to represent to myself, in

addition to the preferences I originally have, the preferences that each party concerned has now or will have in the future. All those preferences become *my own* preferences as to the situations in which I put myself in the positions of others. In this way I can compare the intensity of those preferences, just as I compare my own preferences.

When I ask myself whether I can universalize a certain ought-judgment, I find that the prescription it entails contradicts some of the preferences I have newly acquired by the method of conditional reflection. As previously stated, when I say that I *ought* do X to person P, by the universalizability of ought-judgments this statement must imply that the same thing ought to be done to myself *were I in P's position*. At the same time, by the prescriptivity of ought-judgments it also implies a *prescription* that the same thing be done to myself were I in that situation. Suppose, however, that person P actually prefers not to have X done to himself. When I fully recognize P's position including P's preferences, I should acquire as my own preference one which is equivalent to P's preference, and which can be expressed as 'X should not be done to myself if I were in that position'. However, the prescription implied by my original ought-statement ('X be done to a person similar to P, even if I were in P's position') obviously contradicts the prescription that I have newly acquired ('X not be done to myself were I in P's position').

Then, having such contradicting preferences, how can I make my final ought-judgment? The point here is that all those preferences are my own. Since they are all my preferences, this conflict among preferences can be dealt with as an internal conflict in my mind.

> I can see no reason for not adopting the same solution here as we do in cases where our own preferences conflict with one another[. . .]
>
> [. . .] For [. . .] the interpersonal conflicts, however complex and however many persons are involved, will reduce themselves, given full knowledge of the preferences of others, to intrapersonal ones. And since we are able, in our everyday life, to deal with quite complex intrapersonal conflicts of preferences, I can see no reason why we should not in the same way deal with conflicts of this special sort, which have arisen through our awareness of the preferences of others combined with the requirement that we universalize our moral prescriptions.
>
> (*MT* 6.2 pp. 109–10)

In the case of intrapersonal conflict among my own preferences, if I choose rationally, I will probably make a judgment that would best

satisfy those preferences overall. If I adopt the same method to deal with mutually conflicting preferences in my mind in order to make a final ought-judgment, I will balance those preferences against each other and make a judgment that would best satisfy them overall. Hare assumes that at this point I should be showing a commitment to put equal weight to those preferences according to their intensity. Those preferences are the copies of people's actual preferences, which have been newly formulated in myself by a full recognition of facts, which is necessary for making a rational ought-judgment. Therefore, my final judgment would substantially choose an action that would best satisfy people's preferences after putting equal weight to them according to their intensity. Hare claims that such a decision would virtually correspond with a utilitarian judgment.

Furthermore, when I make a final ought-judgment, I have my own *final* preference, which was adjusted by my newly acquired preferences. This final judgment is the one which I have determined to make, all things considered, even if I were put in any of the positions involved in the actual situation. Thus I am prepared to make the same judgment for all similar situations in which the roles played by particular individuals are exchanged. Hare suggests that this is precisely a universalizable moral judgment which we make to guide an action. We must note here that the 'I', who represents those preferences to himself and balances them to reach an overall ought-judgment, is supposed to be able to detach himself from any of those preferences and to impartially deal with them as if this 'I' is an ideal spectator.

To summarize, Hare's derivation of utilitarianism flows as follows:

1. What are needed for making a moral judgment about what one ought to do are (i) the logic of ought-judgments, (ii) recognition of facts that are relevant to the judgment in question, and (iii) rationality in the sense of aiming to follow logic and recognize facts before making the judgment.
2. Ought-judgments have prescriptivity and universalizability. In short, they are universalizable prescriptions. An ought-judgment shows one's commitment to prescribe the same judgment even if one were put in the position of someone else who would be affected by one's judgment. Therefore, if I am to make a rational ought-judgment, I have to look for a judgment that I am prepared to make were I in someone else's position.
3. The relevant facts that I need to recognize when making a moral judgment are mainly: the present situation in question, the alternative

actions, and the state of affairs that will result from my decision. They include the positions and preferences of others who are actually involved in the situation at issue. In particular, in order for me to ascertain if I can make the same ought-judgment as to the hypothetical situation in which I occupy any of those positions, I must recognize what it is like to be placed in the same position as someone else now occupies. Such a position must be exactly like his own, including his preferences. Therefore, I have to imagine what it is like to put myself in his position with his preferences. During this imaginative process, I must not introduce the knowledge or preferences that I originally had, since I must precisely represent to myself, or to be absorbed in, his actual position with his preferences (the Principle of Absorption).

4. In order for me to say that I know what it is like to be in someone else's position with his preferences, I must now have *my own* preferences, which have equal quality and intensity to his preferences, as to what ought to be done to myself were I in that position (the Conditional Reflection Principle).
5. The final ought-judgment that I would be prepared to make is what I can universally prescribe after imagining the actual preferences of each party concerned, creating in my mind my own preferences which have equal quality and intensity to the ones that the parties concerned actually have. Once I acquire all those preferences, I will treat them as my own. If I choose an action that I ought to do, taking all those preferences into consideration, my choice would be such that will best satisfy them overall. This would virtually correspond to a version of utilitarianism, namely preference-utilitarianism, which claims that the morally right act is the one that will maximize the preference-satisfaction of all parties concerned.

We will notice that in Hare's moral theory there are several key points in addition to his analyses of ought-judgments. Two of them concern the so-called process of 'putting oneself in another's shoes', which is claimed to be necessary in order for one to make a rational ought-judgment: one is the Principle of Absorption, which a person must abide by in order to recognize facts as they actually are; and another is the Conditional Reflection Principle, which a person must satisfy in order for him to say that he has precisely recognized those facts. Furthermore, at the final stage (5) of Hare's argument, it is assumed that when one's own preferences fall into conflict, one can settle it by detaching oneself from all those particular preferences and by making an overall judgment, all things considered.

8.3 Hare's implicit use of Sidgwickean principles

As I have repeatedly emphasized, a remarkable feature of Hare's theory lies in his attempt to develop a utilitarian ethical theory using minimum materials such as the logic of moral judgments and observable (and foreseeable) facts. Another interesting point about Hare's analyses of moral judgment is that the requirement which corresponds to Sidgwick's Principle of Justice is clearly positioned by Hare simply as a *logical requirement* for ought-judgments, and not as a principle that our reflective moral intuition apprehends. Still another noticeable difference between Hare and Sidgwick is that, whereas Sidgwick seldom discussed what role the Principle of Justice plays in the foundations of utilitarianism, Hare claims that the requirement of universalizability plays a major role in establishing his utilitarian ethical theory.

At this point we may wonder whether Hare's theory does not require what corresponds to Sidgwick's Principles of Self-love and Benevolence. If Hare's arguments are perfectly correct, we do not need them. Instead, Hare (1) first clarified that the notion of preference plays a central part in our moral reasoning, by suggesting the prescriptivity of ought-judgments; and second, he (2) adopted the Principle of Absorption and the Conditional Reflection Principle, both of which are needed for the precise recognition of facts, and by doing so he proposed the method of acquiring one's own preferences equivalent to the actual preferences of others or to future preferences. Thus there is no need to introduce additional principles about how to weigh others' or future preferences. The preferences a person considers in making a final moral judgment are all *his own*, and therefore, Hare believes, they will naturally be weighed according to their strength. By logic, a moral judgment must derive from the final preference of the person making that judgment. Hare sticks to this essential feature of moral judgment by adopting the method of intrapersonal comparison of original and acquired preferences, while presenting how a person, who is supposed to be comparing only his own preferences, comes to make an impartial, nonselfish moral judgment – *apparently* without introducing substantial moral principles such as those of Self-Love and Benevolence.

However, despite such skillful tactics, there are at least two problems with Hare's derivation of utilitarianism.[6] One concerns the so-called Conditional Reflection Principle, which requires one to acquire preferences that have equal quality and intensity of others' actual preferences. The other relates to a suspicion that, *after* a person has acquired the copies of others' preferences in their respective positions, he might be

able to put different weights on his original and acquired preferences, during the phase of their overall consideration before making his final judgment.

First, let us examine the Conditional Reflection Principle. Hare's theory must appeal only to the logic of moral reasoning and the observable/foreseeable facts. Therefore, Hare consistently attempts to regard this principle as what is required by the logic of moral reasoning. This is why he sometimes claims that it is 'a conceptual truth, in the sense of 'know' that moral thinking demands' (*MT* 5.3), and sometimes that this requirement comes from the prescriptivity of the term 'I'. These explanations suggest that their truth can be ascertained by our linguistic intuitions about the terms 'know' and 'I'. But is this true? Do we really have to acquire new motives when we say that we 'know' that we would suffer were we in this or that position? Even when we make a statement that we 'know' someone's position at some different time *without* acquiring the exact copy of his preference, that statement would sound meaningful on its own.

Practically, it is impossible for us to *perfectly* imagine other's preferences in the first place. Certainly, the use of 'know' or 'I' in the above context may usually imply that the person uttering it should represent to himself another's preferences *to some extent*. However, it seems to have no implication that such representation should be *perfect*. Logic does not seem to require us to behave *as if* we can perfectly imagine something when we actually cannot. In reply Hare may possibly insist that, though the use of 'know' does not always have this implication, the use of 'know' in a *moral* context requires such a perfect imagining. Nevertheless, the term 'moral' here probably contains additional meanings that were not stated in 8.1 of this book. The present meaning of 'moral' is not confined to an overriding value judgment that can be used as a guide for action.

Even if we put aside the problem of the Conditional Reflection Principle, there is another question about Hare's arguments, that is, the suspicion that we may be able to put different weights on preferences at the stage of their overall consideration before making a final ought-judgment. At the moment of imagining myself being in the position another actually occupies, I certainly must recognize his preferences in that position without any distortion of the facts. We can admit this. Then, provided that the Conditional Reflection Principle is correct, this preference should be reproduced in my mind without distortion. However, the precise reproduction of that preference in my mind does not assure an impartial treatment of it when I balance preferences *after* such an imaginative process.

We should recall the point that we are not immediately moved by any particular preference to make our final moral judgment. Having represented different preferences to myself, I must undertake an overall consideration of them in order to make a moral judgment. At this point, I detach myself from any of those preferences, all of which are my own. Then, when I make a final judgment, it may inevitably frustrate some of my preferences.

Isn't it, then, possible for me to put random or arbitrary weights on my preferences at this stage of detaching myself from all my preferences and, with a cool head, balancing them? For example, why can't I put less weight on my preferences regarding merely hypothetical situations, in view of the position I actually occupy in this real world and the preferences I hold in it? It should be noted that the Principle of Absorption, which requires us not to introduce the preferences that we have in our actual position, does *not* apply to this stage, for it was but a requirement that applies to the stage of recognizing facts about the actual positions of others. This principle does not apply to the *next* stage of detaching ourselves from all the preferences in giving them an overall consideration.

It seems to me that, even when a person deals with an internal conflict of his own preferences in an ordinary sense, he usually makes a prudent judgment not by simply putting equal weight on 'the preferences I would have in different imaginable positions' according to their intensity. He will consider all those preferences *plus* ask which position he does and will actually occupy. It would be quite sound for me to judge that, though I *would* strongly prefer to immediately move out of my present house *if* its roof should be blown off in the near future, I ought not to do so because that is very unlikely to actually happen. Then, why can't I judge that, even if I sincerely admit that I *would* strongly prefer higher wages *were* I in the position of poor employee A, B, C, . . . or Z, their wages ought not to be raised, because I know that I, being a millionaire, am quite unlikely to be actually placed in their positions? To be noted here is that, even if I make an ought-judgment that is virtually favorable to my actual position, I am not showing a logical inconsistency *in so far as I orally affirm that I am prepared* to make the same judgment about all similar situations in which the roles played by particular individuals are different. As long as I show my commitment to make the same judgment about similar situations in which people's positions are exchanged, I am satisfying the requirements of prescriptivity and universalizability. I can even insist that my final judgment is based on the full recognition of relevant facts in Hare's sense, because

I have certainly represented to myself the preferences of others. In short, while having perfect recognition of facts as Hare requires, and while satisfying both logical requirements of ought-judgments, I can still make an ought-judgment that is biased toward my actual position. This is possible because I need to detach myself from all my preferences to make an overall consideration of them before making a final judgment, at which stage I am not bound by the Principle of Absorption.

Sidgwick previously insisted that egoists can also make meaningful ought-judgments (see 4.2.4 of this book). Having admitted this, he pointed out the possibility that the action an egoist rationally prescribes and the one that a utilitarian prescribes can differ in a particular situation. These analyses suggest that making a rational ought-judgment in a situation in which the interests of others are affected, by fully recognizing relevant facts and correctly following the logic of ought-judgments, does not necessarily lead to a utilitarian judgment. An egoist will either (1) not accept the Conditional Reflection Principle as a requirement for the recognition of others' preferences, or, even if he accepts this principle, he will (2) put equal weight only on the preferences that he has in his real life at the stage of overall consideration of all his original and acquired preferences, before he finally reaches an egoistic ought-judgment.

If these suggestions are correct, preference-utilitarianism does not necessarily derive merely from the logic of universalizability and prescriptivity and the recognition of facts. Hare needs an additional rationale for insisting that (1) we ought to accept the Conditional Reflection Principle, and that (2) we should not only treat each position *equally in logic*, but also *put equal weight* on the preferences that we have in each position when detaching ourselves from those preferences to make a final moral judgment. Such claims do not derive from the logical requirements for meaningful ought-judgments, nor from the requirement to recognize observable facts.

However, all these puzzles can be solved if we assume that Hare is introducing the premise that we ought to put equal weight on the preferences of ourselves and others at different points in time, according to their intensity. The claim (2) in the previous paragraph is a substantial ought-statement, that is, that we ought to impartially consider the preferences of ourselves and others at different points in time. If we presuppose this claim (2), then the claim (1) makes sense; for if we ought to put equal weight on people's preferences according to their intensity, it would be quite helpful for us to represent to ourselves, as precisely as possible, their actual preferences in order to estimate their actual

intensity. If Hare insists that claims (1) and (2) are necessary for his ethical theory, he is probably introducing the implicit requirement that one *ought* to weigh others' and future preferences as if they are one's present preferences, putting equal weight on them according to their intensity. However, this obviously corresponds to two of Sidgwick's fundamental principles, which require us to impartially weigh the goods of ourselves and others, and present and future goods, according to their magnitude. This means that Hare's ethical theory cannot be established without assuming additional requirements which correspond to Sidgwick's Principles of Self-Love and Benevolence, and hence that Hare's utilitarianism presupposes these principles in an indiscernible way. If these principles do not derive from the logic of moral judgments nor from relevant facts, we must regard them as introduced by Hare's moral intuition. As Sidgwick declared, the Principles of Self-Love and Benevolence are *substantial* moral principles that are *distinct* from the Principle of Justice.

From the above examination, one significant point of Sidgwick's ethical theory, and the relevance of Independent Interpretation concerning the differences of the three fundamental principles (see 6.2 of this book) become clear. As I predicted, once we understand that Sidgwick's Principles of Self-Love and Benevolence were introduced *independently of* his Principle of Justice, we can clearly see obvious flaws in Hare's argument. Hare presented the requirement of universalizability which corresponds only to Sidgwick's Principle of Justice, and developed his argument for utilitarianism as if it is based only on such logical requirements and the recognition of facts. Despite all this, we find, with proper understanding of Sidgwick's ethics, that Hare's argument uses peculiar tactics which would not have been developed if he did not presuppose additional principles about how to deal with the *quantity* of different 'goods'. We must say that Sidwick's analyses were more accurate than Hare's in presenting not one but three fundamental moral principles, and in pointing out that these principles are to be mutually distinguished by their different ways of treating goods, one being logical and the other two being quantitative.

I still believe that Hare's moral philosophy should remain reputable. Although it contains the flaws just described, Hare's argument excels in its attempt to construct a normative ethical theory in as unbiased a way as possible. I should also add that even though it turned out that we must assume some substantial moral principles to construct a normative ethical theory, it does not necessarily follow that we cannot endorse utilitarianism. If only we can successfully show that the most promising

common basic principle, which each of us can consistently accept and which we can all agree to accept (putting our ordinary moral intuitions aside), is the one that 'we ought to put as equal a weight on others' (and future) preferences as we put on our present ones when we take an impartial point of view', then we can present a coherent utilitarian ethical theory just by adding this principle to Hare's original argument. This utilitarian theory is still simple and clear, using minimum materials such as the logic of moral judgments, observable facts *and* only a few substantial principles that everyone would accept. Nevertheless, it is undeniable that this reveals a challenge that contemporary utilitarianism must face. Even in the most esteemed contemporary arguments such as Hare's, utilitarianism cannot be derived without presupposing the fundamental principle that one ought to put equal weight on one's own and others' (present and future) preferences. This principle does not stem from logic or facts, so we must find some other 'proof' to show the validity of this principle. I must confess I myself cannot present a faultless proof to show the reason why one has to weigh others' preferences as if they are one's own. The best I can say is that, *if we assume* that we ought to take others' preferences into our moral consideration in some way or other, it would be difficult to present a more plausible principle than the one that requires the equal weighing of them, according to their magnitude, regardless of the differences between oneself and others or the differences in time. I will deal with some related issues in Chapter 10 of this book. There I will also discuss the issue of how to measure and compare people's preferences and to reach a moral judgment, provided that we consider the intensity of each individual preference in our moral reasoning.

Another question which is frequently addressed as to Hare's moral theory is whether preference-utilitarianism, which Hare came to advocate thorough the above arguments, is superior to the hedonistic one that Sidgwick supported. I will deal with this topic in the next chapter. There this question will be restated as that of whether the *representation of preferences to ourselves* sways our moral judgments by itself, or whether the consideration of liked or disliked *feelings* is the essential component of determining those judgments.

Additionally, there is a problem concerning Hare's theory, which is that his arguments have no persuasive power for people who never intend to make moral judgments in the first place. Assuming that Hare's arguments are correct, people who are going to make certain ought-judgments about what ought to be done may conclude that they ought to do what rational utilitarians would tell them to do. However, some

people may not need to follow utilitarian judgments. They include those who refuse to make any ought-judgment, declining to consider moral issues or to participate in ethical discussion, and those who do use the term 'ought' but whose judgment is always indifferent, saying 'This is neither the case that I ought to do it, nor the case that I ought not' (*MT* 10.7). Such people are not making any logical or factual mistakes. If they quit making moral judgments altogether and act according to their own selfish desires, we cannot stop these egoistic amoralists through the power of logic. Ways of evading morality and utilitarianism will be discussed in Chapter 11 of this book.

9
A Reappraisal of Hedonism

Another feature of contemporary utilitarianism is the emergence of its preference version in place of hedonistic utilitarianism. According to the preference-satisfaction theory of good, the ultimate end of right actions, or actions that ought to be performed, is the satisfaction of desires or preferences. Thus this theory regards preference-satisfaction as 'the ultimate good' in Sidgwick's sense. What is reasonable for an individual to seek is the satisfaction of his preference, and what one ought to seek is to bring about as much satisfaction of preferences as possible. These preferences may be one's own, may be those of oneself and others, or may be those of all sentient beings, depending on different variations of this theory. It is commonly understood that 'the maximization of preference-satisfaction' means to satisfy as many preferences as possible, to satisfy stronger preferences over weaker ones and to satisfy enduring preferences over transient ones.

The similarities and differences between the preference-satisfaction theory and Sidgwick's hedonism can be stated as follows. First, both make a comparison of different 'goods' by comparing *the intensity of preferences*. Even in Sidgwick's hedonism, a comparison of pleasures is made by measuring the intensity of ideal preferences for those pleasures. Second, the two theories differ; whereas hedonism tells us to consider only a certain kind of preference, namely *preferences for feelings* which the person who feels them has at the time of feeling them, the preference-satisfaction theory makes us also consider various other preferences. Indeed, there are many preferences that are not immediately directed toward feelings. Hedonists only consider 'synchronous' preferences, or the preferences which a person has *at the same time as* he experiences the preferred feeling; but we sometimes have preferences

for a state of affairs that occurs at some different point in time; and the preferred state of affairs is not always pleasure.

Here we should also remember that the satisfaction of preference is not always accompanied by a satisfied *feeling*. Pleasure is a state of consciousness called feeling; but preference-satisfaction is usually understood as the occurrence of a state of affairs in which one's desire is realized (see Brandt 1979, Ch. 7). As a definition of desire-satisfaction, Brandt's explanation would be most lucid: 'Suppose Mr. X at a time t wants an occurrence O at some time t', or at any one of many moments t_i to t_n. Then, if O actually occurs at some one of these times, X's desire has been satisfied. And a greater satisfaction of desire has occurred, if the occurrent O was desired more intensely' (Brandt 1979, Ch. 13, p. 249). As so defined, the satisfaction of a preference does not necessarily imply that a person is gratified by having his preference fulfilled. If the preference-satisfaction theory is to be distinct from hedonism, we should interpret it as seeking to bring about the occurrence of a state of affairs in which a person's preference (whatever it is) is realized, whereas hedonism claims that there is no sense in satisfying a person's preference without generating the desired feeling in that person.

The reason why the preference-satisfaction theory is more popular than hedonism in contemporary utilitarianism is partly because, unlike pleasure and pain, preference is empirically verifiable. Therefore, being favorable for practical use, its notion has come to be adopted by utilitarian philosophers, who were influenced by the same trend in the field of economics (see, for example, Shionoya 1984, p. 388). We cannot directly observe other people's internal feelings, but can relatively easily infer their preferences from their act of choosing or by asking them which alternative they would adopt. Sidgwick also uses preferences to compare pleasures, of course, but hedonism only considers preferences toward pleasant feelings and we cannot discern this limited range of preferences from other preferences just by observing people's choices.

The preference-satisfaction theory has merit not only in practice but also in theory, because it appears to save us from the difficulty of Sidgwick's proof of hedonism. First, it is important to remember here that the concept of preference was interwoven in Sidgwick's definition of pleasure as its indispensable component (see Uchii 1988, pp. 222–4). In addition, the reason why Sidgwick concluded that pleasure is the sole ultimate good was because pleasure is considered to be the only thing that we ultimately *prefer*. Thus even for Sidgwick preference was a crucial concept in the definition of good and the very factor that makes pleasure good; so there seem to be relatively few problems in our shifting from

Sidgwick's preference-based hedonism to the preference theory. Second, we might also concede that the preference-satisfaction theory enables us to dispense with the troublesome 'proof' altogether. Sidgwick attempted to prove that *though we often prefer various other things besides pleasure*, the only thing that we find ultimately desirable is pleasure, that is, the preferable *feeling*. Thus he had to narrow down the candidates for the ultimate good to only one thing, which was pleasure. Still, as Sidgwick himself admitted, it might be difficult to fully explain the desirability of knowledge or an ideal by the pleasure that it brings about. If we talk in terms of preference instead of pleasure, however, we can integrate such apparently nonpleasant yet desirable things into our moral reasoning as factors that also affect our decision-making. There is no need of reducing all desirable things to pleasure, by arguing that the only thing that is ultimately desirable is a kind of feeling that would be preferred (by ourselves or by other sentient beings). We just need to point out that what is ultimately desirable is *what would be preferred in itself*, by ourselves or by other sentient beings. This is almost a tautological truth. Then we will be able to claim that what we ought to seek is the satisfaction of preferences, whether we prefer pleasant feelings or whether we prefer the pursuit of knowledge, ideals, etc. We may also add that the preference-satisfaction theory has no more problems than Sidgwick's hedonism regarding the comparison of various goods. Whatever we prefer, the objects of our preferences are mutually comparable by the common scale of the intensity of those preferences, provided that a comparison of the intensity of preferences is possible, as Sidgwick postulated for the sake of his argument. Therefore, as to the question of whether we can think of any better systematic theory than hedonism to provide a common criterion for comparing goods, we can reply that the preference-satisfaction theory can be such a solution.

The best-known moral philosopher who advocates the preference-satisfaction theory of good is R. M. Hare.[1] In this chapter we will again focus on Hare's theory to examine the preference-satisfaction theory in comparison with Sidgwick's hedonism.

In the following, let us assume for the sake of argument that we temporarily accept all the other devices that Hare used for his argument, the difficulties of which we have suggested in the previous chapter – that is, two logical requirements for moral judgments, the Principle of Absorption and the Conditional Reflection Principle. My point is that, even if we accept all these requirements and principles, we must recognize the prevailing significance of Sidgwick's hedonism when we examine the details of Hare's argument.

9.1 Hare's preference-satisfaction theory

It is evident that Hare considers the position of the preference-satisfaction theory as distinct from hedonism (in Hare's terms, happiness-utilitarianism) in his 1981 book, *Moral Thinking* (*MT*).

> [W]hat I am going to discuss is the interpersonal comparison of *degrees or strengths of preference*, because that is the kind of comparison I need for my own argument. I do not need to discuss anything else but this, because the method we are after turns out to be formulable in those terms, and does not need to mention pleasures or any other kind of utilities.
>
> (*MT* 7.1, pp. 117–8)

> The difference between the two versions [i.e. the hedonistic and preference versions of utilitarianism] will then lie in whether only this restricted class of preferences [i.e. the preferences one has for what happens at that very moment] is considered, or all preferences. That a happiness-utlitarianism can be formulated in terms of the satisfaction of a restricted class of preferences is important; for it enables us to retain the link between it and prescriptions, and thus relate it to our present theory [. . .] But it is still my belief that a full account of the matter would assign weight to *all* preferences.
>
> (*MT* 5.6, pp. 103–4)

Hare supports the preference-satisfaction theory because his analyses of moral judgments and the method of moral reasoning based on those analyses lead him to adopt a preference version of utilitarianism.

The reasons why his moral theory evolves around the concept of preference, rather than pleasure, can be explained by the following three points.

First, as clarified by Hare himself, a person's preference is concerned with his choice of action, and it is also a central concept that determines his moral judgment. The expressions of preferences in language are prescriptions (*MT* 6.1, p. 107), and a prescription is a statement that guides an action. A moral judgment that one ought to do a certain act also has prescriptivity, and the prescription it implies is the expression of the preference of the person making that judgment. When the moral judgment is rationally made, its prescription is the expression of the *final* preference he has come to hold, all things considered.

Second, in order to rationally make a moral judgment, one has to recognize facts, including facts about other people's preferences. Most

moral issues arise when the interests of others are affected; though moral issues are not limited to such cases, Hare believes that we are most concerned with those kinds of cases. A person's interests mean the occurrence of a state which he prefers, or the avoidance of the state which he dislikes. Therefore, a situation in which a moral issue arises usually involves the preferences of others as a matter of fact.

Third, the overall preference of a person making a moral judgment is affected, according to Hare, by the representation of people's preferences to himself. Though my moral judgment is always based on *my* preference, this preference is what I finally hold after an overall consideration of all the preferences I originally had and have newly acquired by imagining, and replicating, people's preferences.

Thus, according to Hare's moral theory, one's and other people's preferences are involved both in the logic of moral judgments and in the relevant facts, as well as in the Conditional Reflection Principle. Hare's preference-utilitarianism is derived from the combination of these three points *plus* the requirement of universalizability of moral judgments. If I am going to make a moral judgment, I have to observe the requirement of universalizability: that is, I must be prepared to make the same judgment for a hypothetical situation in which I am put in the same position as another actually occupies, or will occupy. Therefore I have to vividly imagine what it is like for me to be in another's position with his preferences by precisely observing facts. At this point, if the Conditional Reflection Principle is right, preferences that have the same quality and intensity as he actually has (or will have) must be generated in my mind. Therefore, by choosing an action that will satisfy all my original and acquired preferences as much as possible, I will be virtually choosing an action that aims to satisfy *people's* preferences at various times as far as possible. This action is what I universally prescribe, and therefore it is what I finally judge I ought to do. Thus I make a utilitarian judgment that the morally right action is the one that will maximize the satisfaction of people's preferences at every point in time.

To illustrate this line of reasoning, Hare uses the example of driving a car. Suppose that I am driving a car and estimating how much distance I ought to keep from the car ahead. If I am going to make a rational ought-judgment concerning this situation, I have to consider what I can universally prescribe even if I were put in the same position as the occupant of the vehicle ahead. Thus I must imagine what it is like to be in that person's position as clearly as possible. Especially, I must imagine his position at a time when my car collides with his car, together with his preference for such a collision and the effects of that collision on

his preference-satisfaction. If my imagination were perfect, I would represent to myself my own preferences which would match his in quality and intensity. Then, after balancing multiple preferences in my mind, I would perhaps judge that I ought not to come too close to the car in front of me. This judgment that I ought to perform an act which will best satisfy those multiple preferences is virtually a moral judgment that prescribes an action that would maximize the satisfaction of the preferences of all parties involved, including my original preferences and those of the occupant of the vehicle in front of me. This is how Hare has come to insist that the replication of preferences is the key to moral reasoning, and that it is preference-satisfaction that the right actions should aim for.

9.2 Difficulties

However, it is often claimed that the preference-satisfaction theory has several difficulties, and Hare himself is concerned about them in his *Moral Thinking*. Some people suspect that there might be kinds of preferences which we do not need to, or we *may not*, take into consideration when making moral decisions. Most often regarded as such problematic preferences are so-called external or asynchronic preferences. An external preference is a preference for something *other than* what the person having the preference actually experiences himself. An asynchronous preference is one for what happens at a time *other than* the point when a person has that preference, and this is what Hare calls a 'now-for-then' preference in *MT*.[2]

Should we take external and asynchronic preferences into account? Suppose, for example, that a person has a strong preference that others not be engaged in homosexual acts. Should we put equal weight on his external preference, even if it is beyond his experience that people are engaged in such acts? Or suppose that a youngster used to have a determined preference to become a veterinarian after he graduates from a university, but changed his mind as he grew up. He no longer has that preference at present. Should he take his youthful asynchronic preference into account when determining what occupation he ought to seek now, and should he even make the effort to realize his former ambition? According to Hare's theory, in making an ought-judgment he has to imagine various preferences that would be influenced by his decision. His youthful preference is one of such preferences, for it will be either satisfied or frustrated by his decision at present. When he precisely imagines his past preference, he must now acquire a preference

whose quality and intensity are exactly the same as the past one. If his past preference is stronger than his present preference, he ought to satisfy the past one, since he is now able to realize his youthful dream and he ought to satisfy the stronger preference over the weaker one. However, does this reasoning sound plausible? Ought he really to take his youthful strong preference into account and endeavor to fulfill it? These doubts indicate that, if we are to consider all kinds of preferences, including external and asynchronic ones, we are often compelled to accept a counterintuitive moral judgment as to what we ought to do at present. Of course, a mere fact that it is counterintuitive to take those preferences into account will not offer an effective argument against Hare's theory, since Hare assumes that we must not appeal to moral intuition when we conduct critical moral thinking. Still, it seems that counting these preferences can often lead to an apparently unacceptable conclusion, and not a few people would wonder if this possibility suggests some flaws in Hare's theory. This doubt puzzled Hare himself.

Moreover, when the preference-satisfaction theory takes the utilitarian form, in which one compares the strength of preferences and chooses the act that will bring about the maximum preference-satisfaction, hedonists would propose the following counterargument. Suppose that a person spent most of his life (say, 80 years) as an atheist and strongly wished *never* to have a clergyman at his deathbed, but that he suddenly became weakened when dying and sincerely asked for a clergyman at the last minute. Should we say that we ought *not* to call a clergyman because his strong long-time preference prevails when we compare his preferences at each point in time? Suppose, further, that Mr. X is a volatile person who always has stronger desires for many more things than Mr. Y (assuming that we can compare the preferences of two people) even though the actual satisfaction of X's desire is always less intense than anticipated. Should we equally treat both X's and Y's preferences according to their strengths, and thus always regard X's stronger preferences as having greater weight? (For relevant arguments, see Uchii 1988, pp. 239–40; Brandt 1979, Ch. 7.) Thus the hedonists who oppose the preference-satisfaction theory claim that it is perhaps more reasonable to determine what one ought to do by only considering the strength of happiness that individuals actually feel at the time of their preference-satisfaction.

Being aware of these difficulties of the preference-satisfaction theory, Hare tentatively introduces a 'simplifying assumption' in his *Moral Thinking* to exclude external and asynchronic preferences from his arguments, and proceeds as if his theory is turned, 'in effect', into a

happiness theory (*MT* 5.6, p. 103). Then, without further elaborating on the reasons why these kinds of preferences ought to be excluded, Hare admits that he has left unfinished business as to the treatment of these preferences. He would have had no need for such a reservation if he could doggedly push his logic-and-fact-based argument, defying the apparent counterintuitiveness of considering all types of preferences. Hare's somewhat indecisive attitude shows that Hare himself was embarrassed by the fact that some of those external or asynchronic preferences brought about conclusions which are seemingly unfavorable to his own argument.

Nevertheless, in his *Moral Thinking* Hare clearly stated that it was his 'belief that a full account of the matter would assign weight to *all* preferences', including external and asyncrhonic ones. (*MT* 5.6, pp. 103–4) So, though he was swayed at times, Hare's basic position in *MT* was still a preference-satisfaction theory.

9.3 The Hajdin–Hare debate: Sidgwick's proof revisited

However, a paper that shed light on such debates over the preference-satisfaction theory was issued by Mane Hajdin in 1990 (Hajdin 1990). The intent of this brief paper was to eliminate the aforementioned difficulties of the preference-satisfaction theory while retaining the basic frame of Hare's moral theory. Yet consequently, Hajdin's argument in this paper nudged Hare toward a hedonistic version of utilitarianism.

The point Hajdin noticed was that Hare's own argument would not work if, in putting myself in another's shoes, I only imagined *having* a certain preference while not imagining *the state of mind* that I might have when that preference was satisfied or frustrated. In developing his moral theory, Hare utilized the so-called Conditional Reflection Principle. This principle states that, as a necessary condition for putting oneself in another's position, one must represent to oneself certain preferences, the quality and intensity of which are equal to those of another. Hajdin claims, however, that, if we closely examine Hare's argument as to imagining what it is like to put oneself into another's position, we will soon notice that the crucial step is imagining what it is like for another's preference *to be frustrated or satisfied*.

Hajdin clarifies this point by recalling the previously mentioned example of driving a car. According to him, what Hare actually requires of me in this example, in which I am supposed to imagine what it is like to be the occupant of the vehicle ahead, is to imagine his *experience of collision*.

Hajdin points out that, first, the occupant of the car in front will *always* have the preference not to be crushed by the car in back. In other words, whenever he is asked, he would answer that he would like to avoid the collision. However, Hare does not think that *that* preference is all that I should represent to myself when making an ought-judgment. What I must also imagine is what it is like to be in his position *were the collision to occur*. Thus Hajdin states as follows:

> The important thing to notice is that according to Hare's *own* treatment of this example, the thought-experiment that the method of moral thinking requires me to perform consists not only of my imagining what it is like to be one of the occupants of the vehicle, which includes my imagining what it is like for him to have the preferences that he does, but also of my imagining what it is like for him to have these preferences *frustrated* (or satisfied). That is, I am not supposed to imagine merely what it is like for him to *prefer* not to undergo a collision, but also what it is like for him to *undergo* a collision that he prefers not to undergo (and then presumably to compare that with what it is like for him to have his preference satisfied for not undergoing a collision).
>
> This last element seems to be essential to the force that moral thinking has. It is imagining what it is like for the occupants of the vehicle to undergo a collision that influences me to drive so as to avoid the collision. Merely imagining what it is like to be an occupant of that vehicle while it is traveling smoothly (even with all his preferences) may well leave my driving unaffected.
>
> (Hajdin 1990, p. 306)

So, 'I' must imagine not only the *preference* that the other driver always maintains – the preference not to undergo a collision – but also the *experience* of the collision that he prefers to avoid, that is, the experience at the time when that preference would be *frustrated*. This experience of the collision would mean, more precisely, the *conscious experience* or the *state of mind* that the person undergoing the collision will have.

Here I would like to clarify one point which Hajdin does not explicitly make, but which both Hare and Hajdin seem to be assuming. This point can be explained as follows. First, what I am supposed to represent to myself when I imagine what it is like to be in the position of the occupant of the vehicle ahead would be a sort of *conscious experience* at the time of undergoing the collision. Second, we should recall, however, that there are three different types of consciousness – *cognition*, *volition*

and *feeling* – as Sidgwick correctly articulated. Then, which form of consciousness will play the most crucial part in my imagining what it is like to be in a collision? It would be less relevant for me to envision the other driver's *cognition* of collision (i.e., what he would see or hear at the time of a collision); or to replicate his *will* at that moment – he would most likely not be able to formulate any will because of the shock of the accident. What is essential is to represent to myself the *feeling* that he would have at the time of collision. Of the three types of consciousness, it is not cognition or will but the *feeling* at the time of collision that will most clearly affect my motives and sway my final ought-judgment.

Hajdin also draws our attention to the tone of Chapter 5 of Hare's *Moral Thinking,* in which Hare fully develops his argument on 'the representation of another's preference to oneself'. This chapter is mainly written using such terms as 'suffering' or 'sorrow'; and Hajdin rightly points out that most of its persuasiveness will evaporate if we replace the title of Chapter 5, 'Another's Sorrow', with a flat phrase 'Another's Sorrow-avoiding Preferences'. Sorrow is not merely a preference to avoid suffering but the state of mind in which that preference is frustrated; furthermore, it is not a mere *cognition* or *will* that a person has when his preference is frustrated but the *undesirable feeling* in which his preference is being frustrated.

Thus, if (as Hajdin claims) the crucial step in Hare's reasoning lies not in the representation of a preference to oneself but in the imagining of the *experience* that is desired or undesired, and if the most important aspect of this *experience* is a *feeling*, then the factor that makes the decisive impact on my moral reasoning turns out to be the imagining of the *feeling* that another has when his preference is satisfied or frustrated. The crux of Sidgwick's hedonism lies in this point. What determines our moral reasoning is the *feeling* that is desired or undesired – the *state of mind* that cannot be simply explained as 'preference-satisfaction', which was defined as the *occurrence* of the event that is/was preferred. As I understand it, this is what Sidgwick wanted to claim in his argument of hedonism. If my understanding is correct, in discussing hedonism Sidwick grasped an essential point that Hare missed in his driving example.

My previous argument that it is not cognition or volition but rather *feeling* that affects moral reasoning also shows the significance of Sidgwick's 'proof' of hedonism. Sidwick's proof was not an impeccable one – he simply narrowed down the candidates of the ultimate good by examining each of them while appealing to his own 'philosophical' intuition. However, it has turned out that Hare's preference theory

does not work without referring to *the feelings* at the time when one's preference is being satisfied or frustrated. Now the point is this: Hare has not made an in-depth analysis as to the reason why, in this driving example, the crucial step in my moral reasoning lies in representing such *feelings* to myself. In fact, *feelings* are not the only things that occur at the time of one's preference being satisfied or frustrated. Cognition, volition, physical changes, or effects on one's character will also be involved in that situation. However, in many places in his argument, Hare seems to be assuming that, when making a moral judgment, my choice of actions ultimately depends on the representation of *feelings*, among others. How can he assume this? Hare has not presented any *proof* for the crucial point of his argument that what determines one's moral judgment in this driving example is nothing but *the feeling* that the occupant of the vehicle ahead would have at the time of collision. It was Sidgwick who attempted to provide such proof – proof for the point Hare has overlooked. I will defend this point as an important reappraisal of Sidgwick's ethical theory.

Let us recall Sidgwick's proof of hedonism in the context of Hare's driving example. When considering how much distance I should leave between my car and the car in front, what sways my final moral judgment is not a mere imagining of the preference of the occupant of the vehicle ahead but that of the possible collision in which his preference is being satisfied or frustrated. However, there are many things that may or may not sway my final moral judgment. Will the occurrence of an event that is not related to anyone's consciousness sway my decision? No. When the collision occurs, Uranus may be coming closer to Pluto, but such an event will be totally irrelevant to my moral judgment. Physical circumstances, such as the speed of the vehicle at the time of collision, or the degree of breakage of the vehicle, will be taken into consideration *to the extent that* those circumstances affect my consciousness or the consciousness of the occupant of the car ahead. Then, will I make my moral judgment in this case by considering whether I will carry out a virtuous act or manifest a virtuous character? Probably no. How wise or benevolent I am will depend on my decision as to how much distance I ought to leave from the car in front; and I am about to make this decision. Nor will I decide the proper distance to be taken by considering whether I have a subjectively good will. I will consider whether the victim might be alive or dead after the collision, but the mere knowledge of his survival will not justify my narrowing the distance. What ultimately sways my moral judgment will be the *experience*, or the *consciousness*, that will be had by the person who undergoes the

collision. More specifically, of the three types of consciousness, it will not be his *cognition* (what he sees or hears at the time of collision) or *volition* (what he wills at that moment) but his *feeling* (suffering, pain, etc.) at the time of collision that will sway my decision. This is the feeling that is either desirable or undesirable for the person who undergoes the collision. Thus it is his *desirable or undesirable feeling* that will sway my final moral judgment.

If we proceed this way, our reasoning is essentially the same as Sidgwick's proof of hedonism. The 'ultimate good' means what we ought to aim at; as such, this is what we should consider when making a moral judgment. Therefore, when we admit that what we should ultimately consider when making a moral judgment as to what we ought to pursue is nothing but desirable or undesirable feelings, we are virtually claiming that such desirable feeling is the sole ultimate good to be pursued. If this understanding is correct, Sidgwick's proof of hedonism plays a vital role in fulfilling a previously unnoticed gap in Hare's argument on 'another's sorrow'.

However, this is not the end of the story. In his paper, Hajdin further claims that, if his argument is correct, external or asynchronic preferences will be automatically excluded from moral reasoning of the kind Hare advocated. This claim is supported by Hajdin's following hypothesis, which was obtained through examination of the driving example: One's moral judgment is swayed not by (a) a mere representation of another's preference to oneself but by (b) the representation to oneself of his *state of mind*, in which his preference is satisfied or frustrated.

Let us consider external preferences first. Suppose some person A had a strong preference that others shall not be engaged in a homosexual act, even if he will never know whether this preference is actually satisfied or frustrated. Suppose further that I am about to make a moral judgment regarding such acts, and as a part of that reasoning process I am imagining (a) what it is like to be in A's position with A's preferences, and (b) what it is like for A's preference to be frustrated. According to Hajdin's hypothesis, it is not (a) that sways my moral judgment. However, Hajdin points out that the imagining of (b) will not affect my judgment either. There is no difference between his state of mind in which his preference is satisfied and that in which his preference is frustrated. His preference is satisfied when others are actually not engaged in homosexual acts, and frustrated when they actually are; but, *ex hypotheti*, he never knows whether these others actually performed such acts or not. Therefore, whether his preference is satisfied or frustrated will make no change in his state of mind. If it seems to me that there are

differences between these two states of mind, it is only because I am introducing my own knowledge or evaluation; and such an insertion of my own position is prohibited by the Principle of Absorption (see 8.2.3 of the present book). Therefore, the imagining of (b) cannot have any influence on my decision.

The same is true of asynchronic preferences. Suppose that some person B now (say, at a present time t_1) prefers a certain state of affairs to occur in a future time t_2, but that at t_2 B will no longer prefer it to happen. For example, B, as a youth, now (t_1) has a very strong resolution that music by Marilyn Manson shall be played at his deathbed (t_2), but at the time of his death such a preference will have fled his aged mind and his only hope will be to die quietly surrounded by no one other than his family. Now suppose that, in the process of my moral reasoning about what ought to be done for B at his deathbed, I am imagining (a) what it is like to be in B's present position with B's present asynchronic (now-for-then) preference, and (b) what it is like for this asynchronic preference to be frustrated. However, neither the imagining of (a) nor that of (b) will urge me to play his once favorite music for him at his deathbed. On the one hand, the imagining of putting myself in B's position at time t_1 will not affect my moral judgment, for B's preference at t_1 is neither satisfied nor frustrated at t_1. On the other hand, the imagining of B's youthful preference being frustrated will not motivate me to play Marilyn Manson for him at his deathbed (at time t_2). B's youthful preference is frustrated when the desired event fails to occur when B is dying; but the dying B will have no *frustrated, undesirable state of mind* about it, for he does not, *ex hypotheti*, have the correspondent preference at t_2. Since my moral judgment will not be swayed either by imagining B's asynchronic preference or by imagining that preference being frustrated, I do not need to consider B's asynchronic preference in making a moral judgment.

Hajdin does not present arguments about other types of irrelevant preferences, but by using the same logic, preferences based on a false recognition of facts, or on erroneous beliefs as to the state of mind that a person might have when his preference is being frustrated, will be excluded from moral reasoning – or will be considered less important in moral reasoning. Suppose a person has a strong preference to keep using a credit card even though such expenditures exceed her budget, falsely believing that it will make her feel better. Suppose further that, in reality, doing so only brings her an uneasy, sinking feeling. We can treat such irrational preferences, which are expected to bring no one a desirable *state of mind*, in quite a similar manner as asynchronic preferences.

Getting back to Hajdin's original argument, Hare, who admitted the validity of Hajdin's argument in the driving example, came closer to accept the aforementioned logic to exclude external and asynchronic preferences from moral consideration. Importantly, Hare acknowledged here that his position would be reduced to hedonism if he really accepted the idea that external and asynchronic preferences were to be excluded from moral consideration.

> The effect of the elimination of these two kinds of preferences from moral reasoning would be to make my own theory much more like the 'happiness-utilitarianism' advocated by Richard Brandt [. . .] This is because 'happiness' can be defined for some purposes as a state in which we get the experiences we prefer to have and not those which we prefer not to have.
>
> (Hare 1998, p. 399)

If Hare unreservedly accepts Hajdin's argument, it will turn out that Sidgwick was not only right in having been aware of *a* point that Hare overlooked, or in providing *a* proof to fill the gap in Hare's driving example, but *entirely correct* in defending hedonism.

9.4 Further examination: Is hedonism the whole truth?

Let us ask ourselves, however, whether we should simply return to hedonism. One merit of the preference-satisfaction theory lies in the fact that it enables us to take various preferences, some of which might be irreducible to the notion of pleasure, into moral consideration. Should we abandon this merit and go back to hedonism? We need to stop and carefully reexamine the matter. Hajdin developed his argument by considering only one example – a moral judgment about a distance problem while driving. However, it could be a mistake to generalize a truth about one single case. We should recall Sidgwick's warning that truth must be ascertained by careful reflection. We shall see whether an apparent truth can be overturned by further examination (see 5.1.2 of this book).

In fact, Hare has not completely accepted Hajdin's argument. Though he admitted that he had come closer to accepting that modification of his theory, Hare concluded his response to Hajdin's paper by confessing that he remains uneasy because he is still inclined to take into moral consideration *some* asynchronic nonexperiential preferences, such as the desire about what should happen to his family after his death (Hare 1998, pp. 399, 404).

Hare's uneasiness may perhaps be explained in a way that is consistent with Hajdin's central claim. Those 'apparently external or asynchronic' preferences that Hare wants to take into his moral consideration might be the twisted expression of *non*external or *non*asynchronic preferences. For example, Hare's apparently asynchronic desire that his children should be looked after upon his death might actually be a *moral* preference that he has acquired as the result of his sympathy with his children's preferences. When Hare says that he craves to count his external preferences, he may be virtually insisting that he *ought to* take into consideration his children's *synchronic* preferences to be looked after. By having represented his family's synchronic preferences to himself (in the process of his *moral* reasoning about what *ought to* be done to his family after his death), he might have acquired his own preferences to let them live long and healthy lives. Hajdin similarly claims that most apparently external or asynchronic preferences could be translated as *non*external or *non*asynchronic ones (Hajdin 1990, note 11). If, as Hajdin points out, we can interpret apparently external preferences against homosexuality as 'non-external preferences for the experience of living in a certain type of social order which is inarticulately believed to be threatened by homosexual practices' (*ibid*.) – or, to explain further, as preferences *not to have displeasure* that might be caused by such threats –, these preferences may be taken into consideration when making a moral judgment.

This argument, however, is not conclusive. Here we should ask whether *all* kinds of preferences must be translated as synchronic, nonexternal preferences in order for them to be taken into moral consideration. The crucial question is whether the previously stated assumption, which is that a preference will never motivate us to commit a certain act and sway our moral decisions *unless we imagine the feeling in which this preference is being satisfied or frustrated*, applies to *all* kinds of preferences.

Let me further explain this point. First of all, why does a mere representation of a preference to avoid collision (and to evade the subsequent suffering) *not* sway my moral judgment about how much distance to keep from the vehicle ahead? This is because the mere imagining of retaining such a preference does not motivate me to perform an action, and hence does not affect my decision-making process. This preference to avoid collision is somewhat *latent*; it motivates me and affects my decision making *only when* I imagine how I would feel when this preference is being frustrated, that is, when I experience the collision. This is the reason why the key in this driving example is not the imagining of what it is like to retain a preference to avoid collision but

the imagining of the satisfied or frustrated feeling at the time of collision. Thus the question is whether *every* preference motivates us *if and only if* we imagine the experience or feeling in which this preference is being satisfied or frustrated.

As for me, most of my own preferences fall into this category. For example, the pursuit of knowledge or of an ideal won't stimulate my will to pursue them unless I anticipate the happy and satisfied feelings that will accompany or follow the attainment of such knowledge or ideal. If I knew that I would not obtain any satisfied feeling or sense of fulfillment, I would not be moved to pursue an ideal.

However, we will not be able to conclude that *any* person's *any* kind of preference motivates us only when we imagine the feelings or experiences at the time when the preference in question is being satisfied or frustrated. We know that some people wish a certain thing to happen after their death, or have a strong will as to their way of dying. These people are often motivated to fulfill such wishes (preferences) *even though* they understand that there will be no one who will experience the satisfied feeling when the desired event occurs. We will also admit that some people may well be motivated to pursue a certain ideal no matter what it brings about. They might simply be struck by someone else's disinterested devotion to attain the ideal, and come to form a strong commitment (preference) to pursue it, knowing that it will not bring them any satisfied or fulfilled feeling. Such preferences *do* motivate these people even when they do not anticipate any feelings when these preferences are being satisfied or frustrated. Just by imagining the state of affairs (*not* the feelings) in which what they wish occurs, or just by having such preferences, they are motivated to perform a certain action. If this is the case, then these types of preferences, whether external/asynchronic or not, would have to be considered in making a moral judgment, because these preferences will certainly sway the decision of a person who is making a moral judgment once he represents their preferences to himself and acquires his own preferences that have the same intensity and quality as their preferences. As long as we admit this, we cannot go back to a simple claim of hedonism.[3]

One thing that we can learn from the Hajdin–Hare debate is this: though Hare did not realize this when he wrote *Moral Thinking*, there seem to be *many* preferences whose ultimate targets turn out to be satisfied *feelings* that can be obtained when the desired event occurs. Such preferences do not motivate us if we simply imagine having those preferences; they motivate us only when we imagine the *feelings* at the time of our preferences being satisfied or frustrated. Such preferences, whose

ultimate target is pleasure, will not sway our moral decision, and hence do not have to be considered in making a moral judgment, when it is *impossible* to imagine the feeling that will be obtained. This happens when no one will actually experience this feeling, as in the case of external or asynchronic preferences. What Sidgwick attempted to show in his proof of hedonism was that *all* preferences ultimately target someone's satisfied feelings. This claim sounds convincing to some people, including myself. In my opinion, however, we cannot assert that every person's every preference falls into this category.

Thus my conclusion about Sidgwick's hedonism and Hare's preference-satisfaction theory is as follows. Sidgwick's proof of hedonism might have shown that the ultimate object of *many* preferences is pleasure, or desirable feelings; but his argument is not so complete as to encompass *all* kinds of preferences. Though at the end of his proof of hedonism Sidgwick challenged us by asking whether there can be a coherent theory that can compare different goods in a more systematic way than hedonism, we could respond that a certain kind of preference-satisfaction theory, which has its philosophical basis in R. M. Hare's meta-ethical analysis, will serve that purpose. Thus finally I, myself, still support a preference-satisfaction theory. However, even though I still adopt the preference-satisfaction theory, I would claim that *not* all preferences need to be taken into moral consideration. What we should morally consider is the preferences that will sway our final decision when we represent *them or their fulfillment* to ourselves. What kinds of preferences sway our moral judgment would be ascertained by examining the ultimate targets of those preferences (in other words, by asking what sort of representation would motivate us to perform the act of fulfilling those preferences) in a very similar manner as Sidgwick did when he examined every particular candidate for the ultimate good. In this regard, Sidgwick's effort to provide the 'proof' of hedonism deserves due respect even in the context of contemporary moral philosophy.

10
Interpersonal Comparison and Maximization

In either the preference or the happiness version, utilitarianism is a theory that requires the comparison of people's preferences in making a moral judgment. That is, it seeks to determine what one ought to do by comparing feasible courses of action and by balancing people's preferences regarding the states of affairs that each alternative course of action will bring about.[1] Theoretically, Hare's special version of utilitarianism is designed to dispense with such an interpersonal comparison, since in his theory all the preferences that are balanced are preferences of the person who is making a moral judgment. All of them are his own preferences, including those which he newly acquired by representing others' preferences to himself and which have the same quality and intensity as other people's actual preferences. We will return to Hare's maneuver later. Here, however, let us take utilitarianism to be a theory that requires interpersonal comparisons of preferences, including those of oneself and others. When we pursue this line, we encounter another problem of utilitarianism, which is how to compare people's preferences and integrate them into a moral judgment. We also have the related problem of how to compare a present preference with a future preference. Let us assume, however, that we are discussing both types of problems, that is, the *inter-personal* and the *inter-temporal* comparisons of preferences, when we talk about comparing people's preferences in the argument below.

As Sidgwick rightly pointed out, this type of problem accompanies not only utilitarianism but also many other ethical theories. Anyone who espouses a moral theory that takes people's preferences into consideration must confront the problem of how to evaluate those preferences in his moral reasoning. Even when a person rejects hedonism or the preference-satisfaction theory, as long as he endorses a theory that

considers people's various *goods*, there arises the problem of how to treat those goods and assimilate them into a final moral judgment.

Among all moral theories, however, utilitarianism has a very clear policy about how to deal with people's preferences. Utilitarians attach weight to people's preferences *in proportion to their respective degree of intensity* and conclude that the right action is the one that will maximize the overall satisfaction of people's preferences. To do this, they must assume that the strength of preferences is measurable and interpersonally comparable; and it is for this assumption that utilitarianism is most frequently criticized. It should be remembered here that utilitarianism is not necessarily the only theory that holds this assumption. Any theory that compares the intensity of people's preferences must rely on this measurability-comparability assumption, though such a theory may not be utilitarianism when it compares people's preferences while giving *unequal* weight to preferences that have the same intensity. Yet it is commonly believed that, even if we admit that we have to somehow weigh people's preferences and assimilate them to form a single moral judgment, it is impossible to accurately measure and compare the intensity of such preferences. This criticism of utilitarianism puzzled Sidgwick most, leaving him with a lingering question about the theoretical validity of hedonism.

In this chapter, we will examine the questions of why utilitarians claim to do an interpersonal comparison of the *strength* of preferences, how we can make this comparison, and how such a comparison might lead to the *maximization* of the *sum total* of people's preference- satisfaction.

A good starting point to explicate these questions would be to consider what would happen if we completely *denied* the possibility of such measurement and interpersonal comparison of the strength of preferences. When we abandon this possibility while still attempting to *somehow* consider people's preferences to make a moral judgment about an issue in which multiple parties are involved, one conceivable proposal would be to reach a moral judgment based solely on the consideration of the *ranking* of preferences. It is often claimed that, though we cannot directly know how *strongly* a person prefers one thing to another, we can reasonably tell, by carefully observing his behavior or demeanor, which one he prefers and his *order* of preferences. Thus, if we could derive a moral judgment only by ranking each person's preferences, that would be favorable to those who criticize the measurement and the comparison of the strength of preferences. However, we are faced with a paradoxical situation when we attempt to make a moral judgment without considering the strength of people's preferences. Let us discuss

this paradox by referring to Kenneth J. Arrow's General Possibility Theorem.

10.1 Paradoxical results of rejecting the measurement

10.1.1 Nonutilitarian strategy I: Majority rule and transitivity

Arrow's 'General Possibility Theorem' is an argument that showed the possibility that *any* method, which attempts to derive a social choice by aggregating personal preferences *while excluding the interpersonal comparison of utility*, could generate a situation that we commonly regard as paradoxical (Arrow 1951). Arrow himself discusses this theorem in the context of social decision-making, and especially from the viewpoint of an economist who asks whether we can derive a social welfare function from personal utility functions. However, his argument can be tailored to deal with the question of what kind of moral judgment we can derive from a consideration of people's personal systems of preferences. When his argument is translated like this, Arrow's theorem would suggest the following paradox. Suppose that I am making a moral judgment about what I ought to do. Suppose further that (1) there are more than three states of affairs that I could possibly bring about, that (2) I can determine each person's *preference-ranking* as to those three states of affairs but that (3) I cannot compare the intensity of one person's preferences with those of others. When (1) through (3) hold, and when (4) I am going to take everyone's relevant preferences into account by using majority rule and the rule of transitivity, there is always a possibility that mutually conflicting moral judgments could be equally justified, or that a resulting moral judgment could fall into a vicious circle so that I cannot determine what state of affairs I ought to bring about.

Arrow's original argument is more elaborate, but here I will just present the gist of it by using a simple example. It should also be noted that I will state the points in a somewhat different manner from Arrow himself. Arrow explains his General Possibility Theorem as the argument that public decision-making could be determined by a single dictator's arbitrary will (in an ethical context, this means that a moral judgment could be determined by a single person's personal preference), but I would rather emphasize that such public or moral decision-making could become circular and never reach a decisive conclusion. (I might add that utilitarianism is indeed a theory that could give a greater weight to a single person's extremely strong preference than to all others' preferences. This possibility is often said to be a weakness of utilitarianism, but I do not believe it to be so – it is no wonder that

a strong preference of a single student who is seriously suffering from harassment is to be given priority over all other students' weaker preferences to make fun of him.) Nevertheless, I must admit that the crucial idea of the following discussion contains a hint of Arrow's argument.

We should also note the following. Usually, a single act is expected to cause several possible states of affairs with a certain probability attached to each. Therefore, to be strict, I must examine which alternative courses of action would cause which state of affairs and *with how much probability*. For simplicity's sake, however, in the following argument I will assume that a single act is expected to bring about a single state of affairs.

Now let us consider a simple example. I am contemplating what I ought to do about a situation in which 21 people are involved. There are three possible courses of action that I can take, which are expected to bring about the states of affairs X, Y, and Z, respectively. I am determined to make a moral judgment that can be rightly said to have considered all the preferences of all 21 parties. However, it is assumed that I can never know the *strength* of others' preferences. I can only tell the *ranking* of X, Y and Z in each person's system of preferences.[2] Hence I am going to decide what I ought to bring about by a majority vote and the rule of transitivity. First, when a greater number of people prefer one state of affairs over another, I must judge the former to be morally preferable (I would call this prescriptive judgment 'my moral preference'). Second, if Y is morally preferable to Z and X is morally preferable to Y, then by the rule of transitivity I must judge that X is morally preferable to Z. Moreover, in this case I must regard X as the *most* morally preferable of all possible states of affairs and hence ought to perform an act that is expected to bring about X. Here my judgment that X is *most* morally preferable simply means that X tops the rank of preferred states of affairs, and such a judgment does not involve any evaluation as to *how* preferable X is over Y or Z. Each of the assumptions stated above, (i) a majority vote, (ii) transitivity and (iii) the choice of the highest-ranking state of affairs, seem very natural as the basic procedure to guide a moral judgment without considering the strength of preferences.

However, let us suppose further that, in this specific example, one person P_1 prefers X over Y and Y over Z (this ordering can be expressed as $X > Y > Z$), ten people, P_2 to P_{11}, prefer Z over X and X over Y ($Z > X > Y$) and another ten, P_{12} to P_{21}, prefer Y over Z and Z over X ($Y > Z > X$). (If we include the judgment that someone prefers X over Y *or* is indifferent about X and Y, this can be expressed as $X \geq Y$, but here I will omit this last equation for argument's sake.) Now in our scenario,

since (1) the majority (11 people, P_1 to P_{11}) prefer X over Y and another majority (11 people, P_1 plus P_{12} to P_{21}) prefer Y over Z, I will reason that X is morally preferable to Y and Y is morally preferable to Z. In addition, by transitivity I will further reason that X is morally preferable to Z and conclude that X is the most morally preferable state of affairs and hence I *ought to* bring about X. Here X was chosen over Y or Z, which means that my final moral judgment implies my moral preference for X over Z. *However, this moral preference, X > Z, contradicts the preference of everyone except person* P_1. Because P_2 to P_{21} all prefer Z over X, I should have judged that Z is morally preferable to X. Moreover, (2) when I consider these 20 people's preferences for Z > X together with the majority's preferences for X > Y, by transitivity I could also have reached the conclusion that *I ought to bring about Z* because it now tops the rank of my moral preference. Nevertheless, (3) when I couple 20 people's preferences for Z > X with the majority's preferences for Y > Z, again by transitivity I would reach the moral judgment that I ought to bring about Y. As a result, although I always use majority rule and the rule of transitivity in exactly the same manner, my moral judgment about what I ought to do in this situation becomes quite different depending on my procedure, and this judgment cannot be settled on a single decisive conclusion. If the acts that bring about each state of affairs are mutually exclusive (that is, when I perform an act that brings about X, I cannot simultaneously bring about Y or Z), the three different moral judgments we reached in the argument above must conflict with each other. Or, to put it in another way, I will fall into a vicious circle and cannot reach a decisive conclusion for the following reason. If I reconsider my moral judgment that 'X is the most morally preferable state of affairs' (stated above in scenario (1)) by taking 20 people's preferences for Z > X into account, then I would reach a new conclusion that, since Z is more morally preferable to X, Z is the most morally preferable one. This last judgment, however, can be overturned by taking the majority's preferences for Y > Z into account and reconcluding that Y is the most morally preferable. However, this last conclusion can again be overturned by taking the majority's preferences for X > Y into account and reaching a still new conclusion that X is the most morally preferable. After all, this reasoning can go on endlessly and there is no reason to stop at a certain point and decide what I really ought to do. Such a vicious circle can occur in any situation where ten people in the above example are substituted by number *n* and where the total number of $2n+1$ people are involved (see Figure 10.1).

Why does this paradox occur? One reason is known to lie in the fact that the above example does not describe the *strength* of each

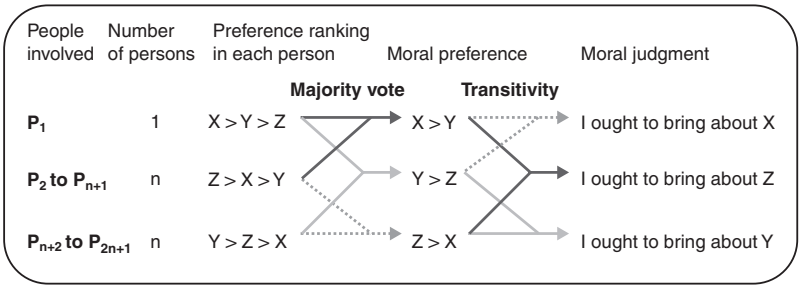

Figure 10.1 The preference-ranking paradox

person's preference (that is, how strongly he prefers one state of affairs to another).[3] Especially in the case described above, we failed to even consider the *greater or lesser gap* (this needs not accurately express the strength of preferences) between the first and second, second and third or first and third places. Note that, when forming the moral preference of X > Y in Figure 10.1, *n*+1 people's preferences of X > Y in the upper two rows (where the gap between X and Y constitutes only one place) are placed on the same level with *n* people's preferences of Y > X in the lowest row, even though the gap between Y and X in the latter preferences takes up two places – Y being the first and X being the third. It is because we adopted such an unequal treatment of the gaps between preference rankings that we have ended up with a circular pattern of moral preferences.

10.1.2 Nonutilitarian strategy II: Scoring according to rank

However, some people may not be convinced by the above argument to accept the measurement and interpersonal comparisons of the strength of preferences. Instead they may propose that, in order to consider the greater or lesser gap between the rankings of preferences, we allot scores to each state of affairs, X, Y or Z, depending on their rank in the order of preferences. Then we tally up the total score for each state of affairs and decide which one is to be chosen. For example, we could give the highest score to the highest ranking one, and lower scores for the lower ranking ones – say, 3 points for first place, 2 for second and 1 for third – and choose a state of affairs that has garnered the highest total score. This strategy is different from the utilitarian method because it automatically allocates fixed points according to the preference ranking, without considering the *strength* of each individual's preferences. When utilitarians give numerical values to preferred states of affairs, they do so

by considering not only the ranking of preferences but also the *strength* of preferences. Therefore, utilitarians may give 3 points to a certain person's weak preference for X over Y and Z, and give 10 points toward another person's stronger preference for Y over X and Z. So in utilitarian thinking, the first ranking one may receive 3, 10 or another number of points, depending on how strongly each individual prefers it.

Now in the nonutilitarian method of scoring according to rank previously stated, there are no worries about a circular conclusion such as we saw in the method of majority rule and transitivity. Nevertheless, this method can generate very odd results. One such situation is that a different moral judgment may result, depending on whether we take into consideration an alternative that will never be chosen in the end.

Let us take an example. Suppose the rankings of 21 people's preferences regarding three states of affairs, X, Y and Z, are as shown in Figure 10.2. I am going to choose the one that is the most morally preferable among all the alternative states of affairs *that I can bring about* (Figure 10.2).

When I can bring about one of the three possible states of affairs (X, Y or Z), I can adopt a method of scoring according to rank, giving 3 points to first place, 2 to second and 1 to third and then tallying up the total score of each state of affairs. Then, the total score of X becomes 33 points (first place in one person's preference ranking, second place in ten people's and third in another ten people's, thus $3 \times 1 + 2 \times 10 + 1 \times 10 = 33$), Y becomes 42 ($2 \times 1 + 1 \times 10 + 3 \times 10 = 42$) and Z will be 51 ($1 \times 1 + 3 \times 10 + 2 \times 10 = 51$); and therefore I can reach a noncircular moral judgment that I ought to bring about Z. We can reach the same conclusion when we generalize this example by giving a points to the first place one, b to the second place one and c to the third place one, provided that $a > b > c$.

However, suppose next that, though there is no change in people's system of preferences, the situation is different only in that I *cannot* realize X. The only state of affairs that I can possibly bring about is either Y or Z,

	1st place		2nd		3rd
1 person	X	>	Y	>	Z
10 people	Z	>	X	>	Y
Another 10 people	Y	>	Z	>	X

Figure 10.2 21 people's preference rankings as to X, Y and Z

and not X. In this scenario, I will naturally consider people's preferences regarding Y and Z only. Thus, assuming that people's systems of preferences do not change from Figure 10.2, we can restate their preferences regarding only Y and Z as shown in Figure 10.3.

If I adopt the same method of scoring according to the ranking in this case, giving 2 points to first place and 1 point to second, my conclusion now becomes that I ought to bring about Y (because the total score of Y is $2 \times 1 + 1 \times 10 + 2 \times 10 = 32$ and that of Z is $1 \times 1 + 2 \times 10 + 1 \times 10 = 31$). We can reach the same conclusion by generalizing this case and giving d points to the first place one and e points to the second place one, provided that $d > e$.

How should we understand the two scenarios stated above? X is an alternative that will never be chosen in the first (Fig. 10.2) or the second (Fig. 10.3) case. It is odd that my moral judgment 'I ought to bring about Z' in the first case changes to the judgment 'I ought to bring about Y' in the second, depending on the possibility of X which is basically irrelevant to my final choice. The idiosyncrasy of this argument becomes clearer when we take a more specific example. Suppose I am an organ-transplant coordinator who is considering which patient, A, B or C, should receive a kidney transplant from a certain organ donor. A is an old man without relatives in a nearby hospital. B and C have ten close relatives respectively, and are in different hospitals distant from the donor's medical facility, C being the furthest away. As for who should receive the transplant, Patient A prefers a ranking of A > B > C, meaning that A should be saved first and that saving B is preferable to saving C. It is natural that A wishes to be saved, and it is also understandable that A believes, if he cannot be saved, B is a more suitable recipient than C because B's hospital is closer to the donor. B and his close relatives (ten people) have the preference of B > A > C. They wish B to be saved first, and believe that, if B cannot be saved, A should be

	1st place		2nd
1 person	Y	>	Z
10 people	Z	>	Y
Another 10 people	Y	>	Z

Figure 10.3 21 people's preference rankings as to Y and Z

saved instead of C because A is closer to the donor. C, who is in a very distant location, and his relatives (ten people) have the preference of C > B > A because they wish C to be saved first and that, if C cannot be saved, B will be saved instead of A (probably because they have a certain sympathy with B and his relatives because their situation is more similar to B than to A). The preferences of all these 21 people should be considered in all the cases in which kidney transplants are performed in their area, for they are all members of the local Kidney Transplant Patient's Network. Now in this case, according to the argument described above, we must conclude that we ought to transplant the kidney to B when it is *possible* to operate on A, and that we ought to transplant it to C when it is *impossible* to operate on A (due to the donated organ's histo-*in*compatibility, for example). This conclusion sounds very strange. Patient A will not receive the transplant in either case. Then why does A's existence sway our judgment as to whether we ought to operate on B or C? One of the reasons for this peculiarity lies in the fact that the strengths of preferences are not being considered. We should note that ten people's preferences for Z > Y in the middle row are compressed when we move from Figure 10.2 to Figure 10.3.

Of course, the mere fact that a certain method leads us to a seemingly odd conclusion does not provide us with sufficient reason to repudiate it. As for the method of majority rule and transitivity, the vicious circle or contradiction occurs only under certain conditions, and some theorists may continue to claim that we can adopt this method for general use by tolerating such extraordinary paradoxes. Nevertheless, we may well want to adopt a more coherent method that will never lead to a circular conclusion or be swayed by irrelevant options, if such exists.

We should admit, however, that from the above argument we cannot immediately conclude that this more coherent method would be the utilitarian way of measuring and comparing preferences. There may still be other theories that do not involve the measuring of the strength of preferences and yet be exempt from any paradox. Post-Arrow theorists of public decision-making have been exploring several such methods. However, it does not seem to be easy to construct a coherent and nonarbitrary method of integrating people's preferences without considering their strengths. I will not discuss the possibility of discovering such nonutilitarian methods in the present work, which focuses on utilitarianism. Instead I will simply point out the following. Once we are allowed to measure the strength of people's preferences on a certain common scale, we can definitely decide the sole action to be done in an nonparadoxical way, by summing up the degrees of strength of

people's preferences toward each alternative and determining the most preferable action. It is John C. Harsanyi (1920–2000) who argues that we can derive a coherent social welfare function from personal utility functions if we are allowed to use cardinal utilities (i.e., the desirability of a preferred object that is expressed in a numerical value proportional to the strength of a preference) (Harsanyi 1976; Harsanyi 1977).

10.2 Key devices for interpersonal comparison: Conversion ratio and extended sympathy

When we do consider the strength of preferences, however, the question is how we *recognize* and *express* other people's systems of preferences, including the various degrees of the intensity of those preferences.

Nowadays, it is being claimed that, as for the system of preferences of a single individual at a certain point in time, we can express it as a scale graduated in progressive degrees of intensity of preferences for various states of affairs, by using the method that was proposed by John von Neumann (1903–57) and Oskar Morgenstern (1902–77). According to this method, we first ascertain the item most preferred by the individual (the top-ranking one) and the item least preferred by the same person (the lowest-ranking one). Once we determine these two poles of this individual's order of preferences, we can then identify the loci of all other items between the two. In other words, we can determine *how much lower* item A is located below the top one and *how much higher* item B is located above the lowest one – by asking the same individual to make a series of choices in which the probability of attaining that goal is assigned to each item. Consequently, all the items preferred by him can be aligned on one scale, from top (most preferred) to bottom (least preferred). We can even give a numerical value to each item, according to its locus on that scale, to express the strength of preference for it, or its cardinal utility. This method is based on at least four unproven assumptions. They include the notions that the relationship between two items in a person's system of preferences (which must show that a person prefers one item over another, or is indifferent between the two) can be explained in a way in which consistency and transitivity are perfectly maintained, and that one can form a preference for a choice to which probability has been assigned. I will not enter into the details of these assumptions in this book. For further information about von Neumann and Morgenstern's measurement of utility and their method of formulating a scale of personal preferences, see Riker 1982, Harsanyi 1976 and Jeffrey 1965 among others.

Now, by using this method, we may indeed be able to obtain the preference scale of a person at a certain point. However, it does not immediately follow that we can compare this scale with the scale of someone else or of the same person at a different point in time (when his tastes have changed). The previously described method only enables us to visualize how various items, more or less preferred *by an individual at the time of his preferring them*, are placed between *his* most and least favorite ones; and the strength of his preference for *his* most favorite item over *his* least favorite one may be different from the strength of another person's preference for *her* most favorite over *her* least favorite. Therefore, even if I heap up different preference scales of various persons at various times, I cannot obtain a common measure to compare those scales. I just collected, so to speak, many measuring tapes whose gradations have different widths. To make an interpersonal comparison of utilities, I have to correlate the scales, in order to make the gradations the same size, and to fix the original point that shows zero utility. The original point would be determined in a way that can be commonly understood, by identifying the point of the status quo, or that of the state neither preferred nor disliked. The key is how to correlate the gradations of different scales.

John C. Harsanyi pointed out and further analyzed the truth that the interpersonal comparison of the strengths of preferences requires this adjustment of gradations. When a person is going to make a moral judgment and hence to compare people's preferences including their strengths, he has to convert the preference scale of each individual at each point of time (including his own) to a scale graded with common units. What he has to do, then, is to set a *conversion ratio* to convert each scale into the scale expressed by the common unit. We can restate this point by expressing, after Harsanyi, Person P_i's preference scale as the personal utility function U_i. Suppose P_i is going to balance the utilities of various states of affairs for various people to make a moral judgment as to which state he should bring about in a situation where *n* people are involved (including himself). Then he has to set conversion ratios $q_1, q_2, \ldots q_i, \ldots q_n$ to convert these people's personal utility functions $U_1, U_2, \ldots U_i, \ldots U_n$ into scales graded with common units. Then he can properly make interpersonal comparisons by comparing $U_1^*, U_2^*, \ldots U_i^*, \ldots U_n^*$, where $U_1^* = q_1U_1$, $U_2^* = q_2U_2$, $U_i^* = q_iU_i, \ldots U_n^* = q_nU_n$. He has to do the same thing for preference scales at various points in time (see Figure 10.4).

In this figure, we are supposing that there are three parties, myself at present (P_i), myself at a certain future point (P_f) and another person at a

certain time (P_o), that will be affected by my moral judgment at present. In making a moral judgment, I have to compare P_f's and P_o's preference scales together with my present preference scale. Each of these scales correctly expresses ***the relative desirability*** of items A to E ***within each person's mind****; but these three scales do not share the same gradations, or units*.

For example, P_i's scale shows that, in my mind at present, the strength of my preference for C over E is four times greater than my preference for D over E. The distance between D and E makes up a unit on P_i's scale. Let us call this unit unit (*i*). Next, the scale for myself at a certain future time (P_f) shows that, in my future mind, the strength of the preference for C over B is five times greater than the preference for E over B. The distance between E and B now makes up the unit for this scale, unit (*f*). Another person P_o's preference scale shows that in his mind the strength of his preference for D over C is three times greater than his preference for E over C (unit (*o*)). However, we cannot say that unit (*i*), unit (*f*) and unit (*o*) represent the same strength. In order for me to compare these

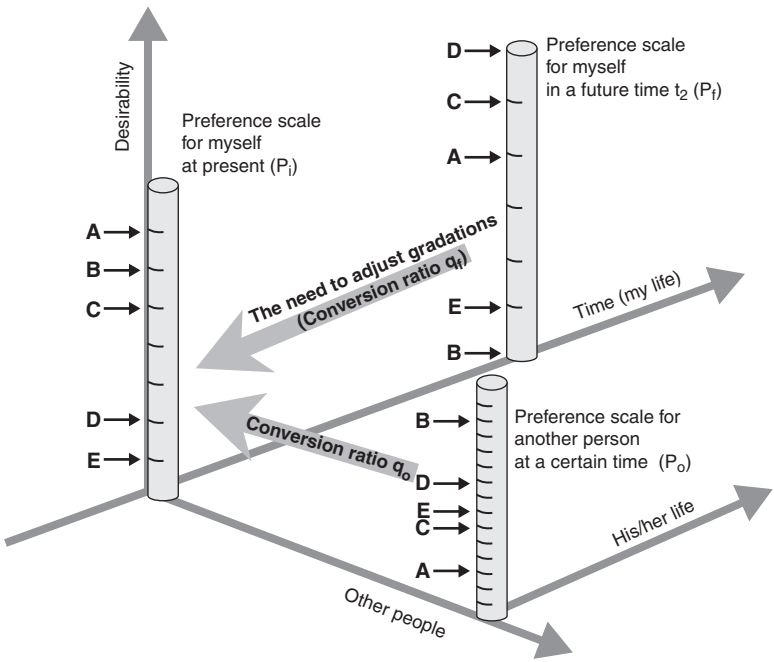

Figure 10.4 Different preference scales graded in different ways

three agents' preferences, I have to convert their scales into scales with standardized gradations, using proper conversion ratios.

Now, it is P_i who sets these conversion ratios, for I am the one who is making the interpersonal comparison of preferences in order to form a moral judgment. Then how do I determine a 'proper' set of conversion ratios? At this stage I have to interpersonally compare the strengths of preferences shown on different scales – or at least the preferences that make up one unit on those scales. That is, I have to interpersonally and inter-temporally compare unit (*i*), unit (*f*) and unit (*o*) in Figure 10.4. How can I do this?

The device that Harsanyi and others have proposed is a kind of thought experiment in which the person making a judgment exchanges his position with others'. Person P_i is to represent other persons' subjective attitudes, including their preferences, to himself, and to form *his own* preference between *being put in the same situation as another person* P_o *with* P_o*'s subjective attitudes* and being in P_i's own position with P_i's own subjective attitudes. This type of preference is what Harsanyi called 'an extended preference'. Arrow, who discussed the previously mentioned paradox that happens when we reject the interpersonal comparison of the strength of preferences in his 1951 book (Arrow 1951), later states the possibility that a similar method could become the basis for interpersonal comparison, by using the term 'extended sympathy' (see the second edition published in 1963, pp. 114–15). The term 'extended' suggests that we need to expand our imagination to form this type of preference. Though our ordinary preferences are about the choice between states both of which can actually occur, this type of extended preference considers at least one alternative for a hypothetical state of affairs, in which we imagine being put in the same position as someone else with the same subjective attitudes as his.

This extended preference enables one to indirectly compare one's own preference with another's within one's own mind.[4] I can indirectly compare P_o's preference for y and my (nonextended) preference for x by representing as precisely as possible P_o's position and his psychological state to myself, in the light of all available facts, and by forming my own extended preference, such as that 'being myself in the state of affairs x is better (in my opinion) than being in the same position as P_o's in the state of affairs y'. Needless to say, this is exactly what Hare attempts to do when he discusses the imaginary exchange of positions and the representation of another's preferences to oneself, which turns interpersonal comparison of preferences into an intra-personal one.

However, the critical point here is that a person who is making a moral judgment is *the only* player in this moral reasoning. It is he who

forms all the extended preferences and sets the conversion ratios. As Harsanyi rightly points out, there is no assurance that my extended preferences correspond exactly with another person's extended preferences. Likewise, there is no guarantee that two different individuals would choose the exact same conversion ratios $q_1, \ldots q_i, \ldots q_n$. My own extended preference would probably differ from someone else's extended preference regarding the same situation: I might prefer to be in the position I actually occupy over being in a position similar to that of P_o, but P_o may form a contrary extended preference for remaining in the position he actually occupies rather than being put into my position. No one knows whose extended preferences 'correctly' represent the strengths of two persons' preferences, one being P_i's preference for the state of affairs x and the other being P_o's preference for y. No one knows whose conversion ratios are the 'genuine' ones.

Representing preferences to oneself, forming extended preferences, and setting the conversion ratios are certainly significant devices for the interpersonal comparison of preferences. These tactics enable one to compare the strengths of preferences or utilities of various individuals *coherently within one's own mind*. However, these methods do not guarantee that the interpersonal comparison in one's own mind agrees with that in another's. If various people's extended preferences or conversion ratios are not exactly the same ones, the preference scales converted within a person's mind and the ones in someone else's could be very different. This means that, even if we agreed on the basic utilitarian policy of giving weight to people's preferences in proportion to their strengths and of choosing the action that most satisfies those preferences, different individuals may still make different moral judgments as to the same situation. Thus there remains the possibility that we cannot attain moral consensus. The biggest problem of the interpersonal comparison of preferences lies in this point. Most of us would wish not only to find a moral judgment that is consistent in one's own mind but to attain a moral judgment that is also acceptable to other people, especially when we are faced with a serious moral issue. The fact that there remains a possibility of never attaining consensus leaves us feeling quite uneasy.

Regarding this, Harsanyi argues that a careful interpersonal comparison of preferences will have an objective validity to a satisfactory degree if we are allowed to adopt the hypothesis that human preferences by and large are governed by basic laws of human psychology. Though at first glance people's preferences may appear to be very different from each other, such differences may be coherently explained by referring to general psychological principles about the effects of various factors

(each person's physiological features, life history, social position, cultural environment, etc.) on human preferences. If this is correct, our judgment about a certain person's preference and its strength becomes more accurate and objectively more valid as our knowledge increases as to the basic laws of human psychology as well as each person's position and his physiological features and circumstances. Richard B. Brandt, who is attempting to compare people's happiness by the strengths of preferences, and R. M. Hare have the same opinion on this point. This claim that basic human psychology is roughly the same among people is but a hypothesis, of course. Still, when justifying the interpersonal comparison of the strength of preferences, we may not be able to expect any better arguments than this.

If we hold such a hypothesis, however, we may not have improved upon the theory of Sidgwick, who realized the need for a fundamental postulate that pleasure and pain are measurable and interpersonally comparable. Theorists after Sidgwick have proposed several devices that enable us to make this comparison, and in that sense we have indeed proceeded a few steps further. Nevertheless, we have not yet overcome Sidgwick in that we still have the problem that is essentially the same as his.

As Sidgwick recognized, the problem of interpersonal comparison is an entrenched one. In my ordinary life, I would continue to compare (or pretend to compare) the preferences or utilities of myself and others as Sidgwick did. If someone were to call for my help when I am about to leave my house, I would decide what I ought to do by comparing the seriousness of my own preference for going out with that of her preference for obtaining my help. My comparison could be erroneous, for my extended preferences and my conversion ratios may be different from someone else's. Still, at times I will decisively make such an interpersonal comparison, in order to make my moral decision. After all, in our real lives we continue to compare different people's preferences, while having the same theoretical problem as that in Sidgwick's time. I must admit that I have left this problem unsolved. All I can say at this moment is that we should be aware of the difficulties in forming moral judgments, all people's preferences considered.

10.3 The maximization of total utility

At any rate, those who tackle a serious moral issue may well venture to compare the strengths of people's preferences to attain a resolute moral judgment. In Harsanyi's view, when we impartially treat various

individuals' utilities expressed by a common scale, we come to support a kind of utilitarianism and choose an action to bring about the maximization of the sum total of those utilities. (Here a utility is meant to be the desirability of an alternative for an individual, with a numerical value being attached to represent the intensity of his preference.) According to Harsanyi, it is an ethical assumption that we are to impartially treat the utilities of various individuals; this means we should impartially treat the strength of their preferences, for utilities have values corresponding to the strength of preferences. Though this is a normative claim, Harsanyi believes this assumption would be commonly accepted by anyone who is making a moral judgment.

Harsanyi further claims that this requirement for impartiality among people will be satisfied when I have no advance knowledge as to whose position I will actually occupy in each state of affairs that could occur. More precisely, this requirement can be met when we assume, in a situation in which n people are involved, that I am not sure whether I will become one or another of those people, and therefore that the probability of my occupying one of those n possible positions is regarded equally as $1/n$. When we assume this, the moral desirability of each possible state of affairs will be expressed by the arithmetic mean of its desirability for each individual. The state of affairs that showed the greatest arithmetic mean would be judged to be the most morally desirable one. Harsanyi claims that, even if we do not attempt such a thought experiment, as long as we maintain the premise that we impartially treat all conceivable utilities, the moral desirability of each state of affairs can be expressed as the sum total, or the arithmetic mean, of its desirability for each person, and the one that shows the greatest number is chosen as the most morally desirable. In any case, the resulting moral judgment becomes a utilitarian one, which aims to bring about the maximization of the sum total of utilities for all parties involved. This utilitarian judgment is coherent. That is, we can always determine the most preferable alternative by defining the moral desirability of each possible state of affairs as the sum total of its utilities for various people, and arranging all possible states of affairs according to their desirability in a transitive and nonparadoxical order.

However, we must give a final word of caution here. We have not yet reached a consensus on the question of whether (1) the impartial consideration of the strength of people's preferences measured by a common scale leads to the weighing of those preferences *in proportion to their strengths* and whether (2) we may measure the moral desirability of each state of affairs by *the sum total* of its desirability for each party involved.

Assumption (1) is what Sidgwick advocated by his Principles of Rational Self-Love and Benevolence and what Hare also presupposed in his moral theory. Claim (2) is assumed in both Sidgwick and Hare's theories as the natural consequence of the treatment of preferences shown in claim (1). However, as John Rawls's theory of justice suggests, there can be another theory in which we give greater weight to the preferences of the least advantaged people rather than giving equal weight to all preferences according to their strength. Furthermore, there can be other methods of tallying up the utilities than utilitarian aggregation. Some may support a Rawlsian difference principle, and others may make a moral evaluation of a certain state of affairs not by the sum total but by the product of people's utilities (see Riker 1982, for example).

I am inclined to accept both (1) and (2) as reasonable assumptions. In a situation where people have conflicting interests, the arrangement most acceptable to all or most seems to be the equal treatment of everyone's preferences according to their strengths. The claim that we should give priority to the weakest people's preferences is touching, but those who are already in advantageous positions may not concur with it. In a heated controversial conflict of people's interests, everyone may claim that he or she is the least advantaged in this situation. The minimum ethical assumption that is acceptable to all these people would be (1) rather than the difference principle.

I am unable to give a definite answer at present to the question of which principle is truly valid – the utilitarian principle of maximizing the sum total of cardinal utilities, the difference principle, *or* the principle of considering the product of utilities. All I can say at this point is that it seems *to me* that the utilitarian principle of summing up people's utilities is most clear and least arbitrary. It would be more appropriate to determine which method of tallying up (or integrating) people's various goods is most effective and plausible, by ascertaining which principle leads to what conclusion when it is applied to practical issues in the real world. To examine this is a future task for me and for utilitarian theorists.

11
Reconciling the Dualism of Practical Reason

The last theoretical problem we will deal with in this book is Sidgwick's dualism of practical reason, or the fundamental conflict between egoism and utilitarianism. According to Sidgwick, utilitarianism and egoism have equally solid theoretical bases; and yet completely different, and possibly conflicting, courses of action can be prescribed by these two different views.

First, from a theoretical point of view, both utilitarianism and egoism are versions of hedonism, or (in a contemporary context) adopt a preference-based value theory. In addition, rational utilitarians and rational egoists are both supposed to admit the truth of Sidgwick's three fundamental principles – the Principles of Justice, Self-Love and Benevolence. Of these three, the Principle of Self-Love constitutes the basis for egoism and also provides *part* of the basis for utilitarianism. Utilitarianism dictates that a person aim for 'people's goods', which are reducible to the good for each individual; thus even in utilitarianism it is assumed to be rational for each person to pursue his own good or his good on the whole as part of people's goods on the whole. Utilitarianism and egoism differ, however, in that the former dictates that a person pursue other people's goods as well. This difference stems from the fact that, unlike egoism, utilitarianism is also committed to the Principle of Rational Benevolence as well as the Principle of Self-Love. Importantly, however, this does not mean that rational egoists *deny* the truth of the Principle of Benevolence. The Principle of Rational Benevolence states what a person is required to do *if he takes an impartial point of view*; but this does not necessarily mean that he *ought to* take such an impartial viewpoint. Therefore, while admitting that one would have to accept utilitarianism *if one took such an impartial viewpoint*, an egoist can continue to pursue his own pleasure alone without any inconsistency, by

proclaiming that he, himself, *never* takes such a viewpoint. As the result, the action recommended by egoism may contradict the action dictated by utilitarianism.

A similar point can be applied to Hare's contemporary version of utilitarianism. An egoist may admit that, *if* one is determined to make a moral judgment that satisfies the conditions Hare sets (including the ones he implicitly assumes), the resulting judgment would certainly be a utilitarian one. Still he can dodge a utilitarian conclusion without any inconsistencies by refraining from making such a moral judgment. Moreover, according to our previous examination in Chapter 8, an egoist can even claim that his final judgment, 'I, Person A, ought to bring about A's greatest satisfaction, and A's satisfaction alone', perfectly satisfies Hare's logical and factual requirements, including universalizability. This is possible if this egoist does not apply the Conditional Reflection Principle to any other experiences *than the ones that occur within his life*, or if he gives greater weight to his own (present and future) experiences when considering all people's preferences to make a final ought-judgment.

Thus utilitarianism and egoism are both theoretically consistent and well founded, and yet in practice they could conflict with each other. Then how can we reconcile them? Could we formulate a convincing argument that the ethical position we ought to take is *not* egoism *but* utilitatianism?

11.1 Some attempts

Derek Parfit (1942–) has presented a unique argument to challenge Sidgwick's puzzle that both utilitarianism and egoism can exist as two independent, and potentially conflicting, theories (Parfit 1986, Part III). Based on his own theory of personal identity, Parfit questions the real significance of the distinctions among people. According to him, there are no substantial grounds for rigorously distinguishing myself from all others, all of whom are commonly believed to exist over time. Nor is this distinction always clear. Based on recent scientific findings (such as the advanced transplant techniques in today's medicine, the knowledge of a human body and brain, and the scientific study of the relationship between mind and brain), Parfit claims that whether a certain conscious body is mine or someone else's can sometimes be blurred. For example, when a neurosurgeon gradually replaces 1, 2, 3 . . . or even a 100 per cent of my brain cells with someone else's, nobody (including me at each moment during the surgery) will be able to find the exact

point where I quit being myself and become another person. As Parfit tactfully demonstrates, the essence of 'me' simply consists in particular experiences, such as feelings, desires, and memories that make up a continuous stream over time. We do not have to assume the continuous existence of a certain distinctive entity, or substance, that can always be clearly distinguished from all other persons. Our body cells and physical appearance greatly change over time. Psychologically, we forget part or most of the feelings, desires and memories that our past selves had, and we often do not know what feelings, desires and memories our future selves will acquire or retain. In a sense, our past or future selves are similar to others. Parfit claims that, if this view is correct, an individual has no reason to confine his concern to a particular stream of experiences that is commonly called 'his own life' (For a more detailed argument of the topic, see Parfit 1986; Nakano-Okuno1997 and 1998b). There are no fundamental differences between my consideration of certain future desires or feelings that are usually called *my future* desires and feelings, and my consideration of certain other (present or future) desires and feelings that are usually called *another's* desires and feelings. Rather, in Parfit's opinion, the viewpoint which my present self takes when I consider my future feelings is *essentially the same* as the impartial viewpoint I take when I consider another's experiences. We do not have to distinguish the Principle of Rational Self-Love and that of Rational Benevolence. If Parfit is correct, there is no reason for egoists to disown the Principle of Benevolence while positively endorsing the Principle of Rational Self-Love.

I believe that Parfit's argument is valid and effective against egoists to a certain extent. If an egoist's claim is based on the belief that 'I am essentially a continuous entity or substance clearly distinct from others, and therefore my life is all that matters to me', then by admitting Parfit's theory of personal identity he would lose the grounds for exclusively supporting egoism. However, Parfit's argument does not have the power to convert all egoists to utilitarians. I may value my own life higher than other people's lives not because there exists a special distinct entity called myself, but because my present self feels a special attachment to the particular stream of experiences that is usually called my life. If this is the case, Parfit cannot criticize me for demonstrating an erroneous understanding of persons and personal identity. There may be no reason to confine my concern to my own life, but there is no reason *to forbid* maintaining a special affinity for my own future self. Sidgwick himself suggests that a human life might be regarded as a stream of various feelings, memories, desires and other experiences; and he further

states that if we take this view, it is philosophically questionable why our present selves must be concerned more about our own future feelings than about other people's feelings (*ME*7 pp. 418–19). Yet Sidgwick observes the fact that we commonly distinguish our own lives from others', regard this distinction as quite important, and have special affinities for our own lives. Even if there are no philosophical grounds for distinguishing the Principles of Self-Love and Benevolence, likewise there is no reason not to distinguish them. This being the case, we cannot refute egoists who give priority to their own lives based on this common-sense distinction. Sidgwick thus concludes that the dualism of egoism and utilitarianism must inevitably remain (see *ME* p. 498). Parfit's argument is not conclusive to solve this dualism.

Meanwhile, when our impression that a human being has a natural attachment to his own life is intensified, we might come to believe that we are inevitably egoistic by nature, and that utilitarianism should be based on the egoistic nature of humans, if it is to be proved at all. One such argument, the proof of utilitarianism based on psychological egoism, was briefly discussed in Chapter 7 of this book (7.2.3). One version of such a proof argues that a human, whose nature is basically egoistic, naturally supports ethical egoism as his fundamental policy, but practically adopts (or pretends to adopt) utilitarianism in order to pursue his egoistic goals. However, we cannot accept this line of argument. What Sidgwick repeatedly emphasized were the following two points: though almost everyone is certainly interested in his own happiness, (1) one does not always exclusively seek one's own pleasure (the denial of psychological egoistic hedonism), and (2) what one regards as the ultimate good to be pursued is not necessarily limited to one's own pleasure, but can only be identified as *someone's* happiness or the pleasure of *some* sentient being (the proof of the hedonistic value theory *in general*). The same points hold, *mutatis mutandis*, even if we adopt the preference-satisfaction theory instead of hedonism. Then the above statement would be paraphrased as (1) a person does not always exclusively seek the satisfaction of his own preferences which were formed prior to moral thinking, and (2) what he judges to be preferable is not necessarily limited to the satisfaction of his own premoral preferences. (It is, however, tautological to say that one always seeks the satisfaction of one's own preference in its broad sense, if we include one's moral or benevolent preferences into the notion of 'one's own preferences'.) As far as these statements are true, to base utilitarianism on the egoistic nature of human beings misses Sidgwick's main point. We are sometimes selfish, but sometimes not.

11.2 Brandt's approach

Another interesting approach to the proof of the supremacy of utilitarianism is Richard B. Brandt's double-edged argument (Brandt 1979, especially Part II). This is an argument that, *whether egoistic or benevolent*, humans would unanimously adopt utilitarian *public morality* as long as they are *rational*.

The logic is plain. We rejected the egoism-based proof of utilitarianism (i.e., the argument that all of us, being essentially egoistic, would eventually adopt utilitarian ethics out of prudence) by claiming that we humans are *not always* egoistic. Then, how about claiming that all of us, *whether psychologically/ethically egoistic or not*, would ultimately think it rational to become, or pretend to become, utilitarians? To put it in general terms, the following logic is sound: 'If A, then B; if C, then B; it is *either* A *or* C; therefore, B always holds true'. If benevolent and egoistic persons could both be persuaded to adopt utilitarianism in practice, we may simply conclude that utilitarianism is the sole practical position that will ultimately be upheld. Then the apparent conflicts between utilitarianism and egoism will eventually evaporate and not disturb us any more.

In the next section, I will examine Brandt's case for utilitarianism in more detail. We should note here, however, that what Brandt provides us is the proof of utilitarianism *as public morality*. He attempts to demonstrate that humans would, *if they were fully rational*, unanimously adopt the utilitarian *social moral system or social moral code*. Thus we must first understand what Brandt means by 'rational'. We also need to clarify his term 'social moral system/code'. Also to be remembered is that the following is a somewhat paraphrased explanation of Brandt's argument. Although I believe my understanding of Brandt is basically correct, my explanation of his theory is considerably simplified, and my own interpretations are intertwined in several places.

11.2.1 *Social moral system* and *rational* choice

Just like Sidgwick and Hare, Brandt endeavors to discern the guiding principles for our moral decision-making, by appealing to as few moral intuitions as possible. Indeed, he proceeds with even fewer assumptions than Hare embraced. While Hare's argument centered around the question of which moral theory would be supported by a person *who is about to make a moral 'ought' judgment*, Brandt considers the question of which moral system, if any, a person would support if he were *rational*.

Brandt uses the term 'rational' to refer to 'actions, desires, or moral systems which survive maximal criticism and correction by facts and

logic' (Brandt 1979, p. 10). His notion of rationality is, like Hare's (see 8.1 of the present book. Hare himself insists that he adopts Brandt's very definition of rationality), looser than Sidgwick's notion of rationality, which meant the intuitive grasp of a certain truth. In my opinion, Brandt's definition is even looser than Hare's notion of rationality. According to Brandt, 'a moral system that would be supported by a person who has undergone the maximal criticism and correction by facts and logic' is the one that would be supported by him if he vividly envisioned and repeatedly reflected on all factual information that is available to him, and if he conducted logical reasoning. 'Logical reasoning' here merely means the reasoning that makes sense to us, in the normal meaning of the term 'makes sense'. It does not particularly mean to follow the logic of moral judgment as Hare insisted, and it does not even imply that everyone must strictly follow the principles of logic. For Brandt, the moral system that a person rationally supports is the one that *he* would voluntarily choose to support after *he* has carefully considered choices based on facts and logic.

Brandt gives this minimal meaning to the notion of rationality for the following reasons. Whether we have grasped a so-called truth or not, it is certain that we often ask, and seek to give an answer to, the question of what (actions, desires, moral systems, etc.) we would choose or support on careful reflection. It is important for us to make a well-considered decision. Now, to make a well-considered decision, one should consider various facts in a 'logical' manner, that is, in a way that makes sense to us. Thus Brandt used a single term 'rational' to describe this process of considering facts and logic, and decided not to add extra meanings to it (Brandt 1979, Part I, Ch. 1. For details about Brandt's notion of rationality and its significance, see Nakano-Okuno 1998c).

One assumption in Brandt's argument is to be noted here, however. He presumes that, when a person targets a certain purpose (or when he has a certain preference), he would surely adopt a means suitable to that end (or to the realization of the preferred state of affairs). Brandt does not regard this assumption as part of the *meaning* of the term 'rational'. He simply thinks that it is a pattern of human psychology as described in theories of psychology. According to Brandt, to choose a suitable means to a designated end is what a person would almost certainly do if he considered it based on facts and logic.

'A moral system/code' is a system or code which controls our behavior, but which differs from statutory laws and other legislation (Brandt 1979, p. 163 ff.). The term 'system' suggests a somewhat complex structure that functions according to a certain definite principle, or a set of

definite principles that can be called 'a moral code'; or, if such principles are unknown, a structure that has certain features and somehow functions in a well-ordered manner. Brandt, himself, does not clearly define these terms, but he states that a person has a moral system/code when he embodies several features as explained below.

A moral system may be supported by an individual as his own personal policy, or as what should be *current* in a given society. ***A personal moral code*** means a person's internal surveillance system for behavior, which is commonly called 'conscience' in a broad sense. It is normally accompanied by certain motives for acts, and certain feelings related to performing or not performing such acts – comfort, remorse, approval/disapproval, and so on. Each individual believes that what his personal moral code prescribes should be justified, and the prescribed act can be expressed in such terms as 'the right thing to do'. While each person has his own personal moral system, the actions that are prescribed by such a system can differ considerably from person to person. (Plus, there are a few who have no personal moral system.) Here we may consider an egoistic policy as one of the personal moral codes, for such a policy can be expressed by the principle that 'I *ought to* maximize my own happiness, or preference-satisfaction, throughout my life'. This would expand the meaning of the term 'morality' as commonly understood; however, as Sidgwick pointed out, if a person believes that it is his legitimate duty to follow this egoistic principle, then it can be properly called his *personal moral system* in Brandt's sense. We may also call it his *personal ethical view*.

Most societies also have social moral systems. ***A social moral system*** is the system to control behavior which is or should be prevalent in a given society, as applied to people that belong to that society. When this system is guided by a set of definite principles, we call these principles 'a social moral code'. However, a social moral system is not necessarily guided by a limited number of principles.

Such a social moral system is not one stipulated by a government or legislation. Rather, it becomes *current* in a society by being supported by most or all of the people in it. Basically, each individual is free to choose whether to endorse a certain social moral system. One may choose to maintain the social moral system that is already current in society, or may decide to support a completely new one. One can even choose to endorse none. For an individual to endorse a certain social moral system is for him (1) to agree that a certain system of controlling people's behavior become widely accepted in his society, (2) to decide that he, himself, will publicly follow (or at least pretend to follow) its dictates, (3) to agree

that any conflict of interest among people is to be arbitrated by appealing to its code, and (4) to agree to use such expressions as 'morally right' to denote the actions that this system prescribes. When a person endorses and thus agrees to abide by a certain social moral system, he is indirectly yet voluntarily deciding what he morally ought to do and is motivated to carry it out. However, it is not always true that he feels comfortable in obeying the mandates of the social moral system he endorses, and is gnawed by a sense of guilt when he breaches them. He may endorse a certain social moral system and agree that it should be widely accepted by people including himself, but the specifics of that system may not perfectly coincide with his own personal moral system – or he may not have any personal moral system. Still, it is proper to say that a social moral system comprises norms for each individual's acts; and, unlike legislation or politics, it is the individual who determines whether to endorse a certain social moral system. In this respect, the consideration of social moral systems also belongs to the realm of ethics.

Brandt's main interest lies in whether a person endorses a utilitarian *social moral system* that should be prevalent in his society, rather than whether this person honestly adopts utilitarianism as his personal moral view. Brandt's main question can therefore be restated as follows: considering all available facts and logic, will a person prefer to have a certain social moral system be prevalent in society rather than none, and if he does, which social moral system will he endorse?

This is a matter of choice among possible alternatives. Generally, a person makes a considered choice by examining alternatives, their feasibility, and his own preferences. So our decision depends on alternatives among conceivable social moral systems, the feasibility and sustainability of each alternative, and our preferences as to whether to endorse a certain social moral system.

We can conceive of numerous variations of social moral systems, but obviously infeasible ones can be excluded from our consideration. It is useless to support a moral system which we cannot implement or which cannot become prevalent and persist in society, for we support a social moral system (if any) in order to publicly control our actual behavior. In fact, some types of social moral systems are presumably impracticable. For example, a malicious social moral system will not persist long because of its devastating effects. One that requires people to strictly carry out definite duties without exception also seems to be infeasible, for we often question why we ought to obey rules, especially when our interests conflict with each other, and we sometimes find it difficult to judge which rule to obey in a particular situation.

We also need to identify types of preferences that are relevant to choosing among moral systems. Those will include preferences for certain features of a moral system or code (such as kinds of prescribed actions or binding forces of that system) and preferences for certain consequences that will result when a certain system is maintained and widely accepted. Brandt enumerates the following four types of preferences. (See Brandt 1979, p. 203 ff. I reclassified and renamed Brandt's original list of what he calls 'valenced outcome'.)

(i) ***Intuitive desires or aversions toward certain types of dispositions or actions of members of one's society***. Examples of these include cases in which one intuitively prefers that people be compassionate, or intuitively dislikes gambling.
(ii) ***'Egoistic' preferences for one's own happiness or satisfaction.***
(iii) ***'Benevolent' preferences for people's happiness or satisfaction.***
(iv) ***Other moral or intuitive preferences***, such as those for economic and social equality or for reward in accordance with merit, etc.

In addition to (ii) and (iii), there can also be other types of preferences, such as those for the happiness and satisfaction of one's relatives or loved ones. Such preferences, however, can be regarded as a variation of egoistic/self-regarding preferences in that they are matters about which one is personally concerned, or as a variation of benevolent preferences whose target is limited to what a person can actually realize. (Brandt himself seems to regard this type as an imperfect form of benevolent preferences.) Furthermore, benevolent preferences usually refer to the ones which seek people's happiness or satisfaction including one's own, but there are some people who wish for the happiness of others while neglecting their own interests. Strictly speaking, such preferences are not the same as benevolent ones as previously explained, but I will consider them as variants of benevolent preferences.

Brandt argues that preferences (i) and (iv) can be excluded from our consideration in making a rational choice of a moral system. This part of Brandt's argument exactly corresponds to Sidgwick's proof of hedonism. On the one hand, intuitive preferences as to people's dispositions or actions, (i), seem meaningless unless we hold those preferences out of further motives such as (ii) or (iii). It also seems that preferences of type (iv) actually originate in benevolent preferences (iii), and the limits and exceptions of preferred equality, rewards, etc. are explicitly or implicitly determined by such benevolent consideration. On the other hand, after analyzing egoistic preferences for one's own happiness or satisfaction

(ii) and benevolent preferences for people's happiness or satisfaction (iii), Brandt concludes that these two types of preferences are based on no further motives and are irreplaceable with other types of preferences. Thus he attempts to discuss the rational choice of a moral system by considering what a person would choose if he had preferences of type (ii) *or* (iii). Like Sidgwick, Brandt claims that we humans have benevolent desires as well as selfish ones, and that these types of desires can both be regarded as rational and can seldom be reduced to other types of preferences. To put it in Sidgwick's words, various targets of preferences *other than* happiness or satisfaction (such as those for inanimate objects, virtues and duties) will, upon reflection, be regarded as not desirable in themselves; however, happiness or preference-satisfaction will continue to be regarded as 'desirable' even after we repeatedly consider their true values. Here Brandt is accurately grasping the following points of Sidgwick's argument for hedonism: (1) a person does not always pursue his own pleasure *alone* (the denial of psychological egoistic hedonism); and (2) what a person regards as the ultimate good can be identified as the pleasures of *some* sentient being, which are not necessarily limited to his own pleasure (the proof of hedonistic value theory in general). Thus Brandt proceeds along Sidgwick's line of argument, and yet attempts to support utilitarianism as public morality.

Now, from the discussion above, we will focus on two types of preferences, egoistic and benevolent, as the preferences that are relevant to the rational choice for a moral system. Then, our question becomes: (1) 'which social moral system would a person support, if any, when he has *egoistic* preferences?'; and (2) 'which social moral system would a person support, if any, when he has *benevolent* preferences?'

Brandt points out that, in reality, some people may have a stronger tendency to be benevolent and others may be more inclined to be egoistic. Some may have no iota of benevolence even after considering all available facts and logic. According to Brandt, benevolence in a person is fostered by experiencing sympathetic ties between another's pleasure/pain and those of his own in his early childhood. (Thus benevolence is an acquired disposition; however, since it is usually ingrained in early childhood and afterwards reinforced by warm relationships with other people, we normally have benevolent preferences to a greater or lesser extent and consider such preferences to be desirable ones.) This being the case, benevolence may not develop in a person's mind if he was abandoned as a child and/or grew up in an environment in which it was hard for him to associate other people's delights or agonies with his own. Benevolence can be hindered from developing in a person if

his first attempt to help someone was criticized or rejected, or if people surrounding him were unkind and violent. It may be difficult, if not impossible, for a person who grew up in such an environment to come to believe, on reflection, that benevolence is a desirable tendency to possess (Brandt 1979, Part I, Ch. 7). In contrast, perfect benevolence means that a person impartially wishes all people's happiness or satisfaction equally as much as his own happiness or satisfaction. Many or most people will lie in-between these two poles; we will experience times when a benevolent desire surges and when an egoistic desire prevails. In any case, our rational preferences – that is, the preferences that we will continue to have even after undertaking the process of maximal criticism and correction by facts and logic – can be categorized as either of the two types, egoistic or benevolent. Then, a consideration of the questions presented in the previous paragraph, (1) and (2), enables us to identify a social moral system that we will rationally support, that is, a social moral system which we would support after careful reflection based on facts and logic.

11.2.2 The double-edged argument for utilitarianism

First, let us consider which social moral system a person would rationally choose when he has benevolent preferences. Brandt insists that this case would obviously support a utilitarian moral system. When a person whose predominant preferences are benevolent ones is presented with two alternatives to make people happy and satisfied, he would, if he were fully rational, choose the one that will bring about the greater sum total of people's interests, after giving equal weight to everyone's interests. (Hereafter I will use the term 'interests' to express happiness or preference-satisfaction of the kind I previously described. In addition, like Sidgwick and Hare, Brandt holds a simple view that the aggregation of people's interests would be expressed as the *sum total* of them.) When a person is motivated mainly by such benevolent preferences, he will support utilitarianism as his *personal ethical view*. Furthermore, he would endorse a *social moral system* in which each person seeks to maximize people's interests. This is because people's interests would be systematically and steadily promoted if such a moral system were widely adopted. Thus this benevolent person supports utilitarianism as the guiding principle for his social moral system. Of course, this means that he himself agrees to act according to this utilitarian social moral system as the 'morally right' thing to do.

However, although this social moral system has a utilitarian principle at its core, it does not necessarily order an individual to calculate

people's overall interests and to endeavor to maximize them in each particular situation. Rather, this utilitarian social moral system will guide us to generally act according to what Sidgwick called middle axioms, or the rules that would promote people's interests in the long run if obeyed by a large number of people. Such rules will include 'do not kill', 'do not steal', 'keep your promises', and other maxims that we are quite familiar with. This is because, as Sidgwick clearly stated (see 7.4 of the present book), it is difficult and sometimes impossible for us humans to accurately calculate people's interests on each occasion and it would therefore be expedient for us to usually follow general rules that are carefully chosen. By voluntarily following such rules, however, an individual is taking a voluntary action to bring about the maximal interests of people in the long run; in this regard he is certainly acting according to the utilitarian principle of public morality. By so acting, this individual demonstrates that he holds the view that it is morally right to act in this 'indirectly utilitarian' way, at least in public.

Next, let us consider which social moral system a person would rationally choose when he has egoistic preferences. He will attempt to maximize his own interests. Brandt claims that, in this scenario, he will still wish for the prevalence of a certain social moral system rather than none, unless he is in an especially privileged position. Even from an egoistic point of view, he would favor a certain social moral system because it would control other people's behavior and protect him from their attacks and vilification. This will also enable him to predict other people's behavior, which is more or less guided by general rules. In addition, most of us may well appreciate, even from egoistic motives, the value of mutual trust and cooperation that will be secured by having and maintaining a commonly accepted moral system, for such trust and cooperation will give us peace of mind and enable us to attain tasks that are difficult to achieve when working alone. A person with egoistic preferences will certainly support a social moral system that is expected to bring about these consequences.

Which exact moral system will a rational egoist support, however? When you have egoistic preferences, you will naturally be inclined to adopt ethical egoism *as your personal ethical view*. On the other hand, you cannot endorse an egoistic *social moral system*, which requires *each person* to maximize *his or her* own self-interests. If everyone in society is allowed to freely pursue his or her own interests, serious conflicts among people would occur, and, unless you have the method to arbitrate these conflicts, it is unlikely that you can attain the maximum personal happiness or satisfaction that you, yourself, wish to obtain.

However, it will also be unreasonable for you to endorse a social moral system that requires *everyone else* to promote *your* interests, for such a system will presumably be unfeasible. If others are equally rational and egoistic, a system that is favorable only to yourself will never become prevalent, so you will not be able to attain any of your personal interests by overtly supporting it. According to Brandt, if you are trying to find a *feasible* social moral system for a society, which is full of equally rational and egoistic persons, you will soon notice that you, yourself, have to accept the same terms that you are requesting from others. The social moral system that you can rationally support and accept based on purely egoistic motives will be one that requires each member of the society to equally respect the interests of all people including himself and others, and to collectively bring about the maximum surplus of people's benefits (happiness or satisfaction) over burdens (unhappiness or dissatisfaction). Brandt believes that this is the system governed by a utilitarian principle. Thus, according to Brandt, even when a person has selfish preferences, it is rational for him to endorse utilitarianism as a social moral system. Such an egoistic person will not be disinterestedly obeying this utilitarian social moral system, and will not be wholeheartedly dedicated to the (direct or indirect) promotion of people's interests. However, he will agree, at least in public, to act according to what this utilitarian moral system orders him to do. He will also agree, at least in public, to decide social issues by appealing to utilitarian thinking.

11.3 Unsolved problems

If Brandt's double-edged argument is correct, everyone would support a kind of utilitarian social moral system were he fully rational, if he is benevolent or selfish by nature, and whether or not he adopts egoism as his own personal ethical view. Even if we suppose he does not adopt a total maximization principle, he will endorse a social moral code that somehow requires benevolence, or equal consideration of people's interests. Thus we may conclude that utilitarianism, or at least some form of benevolent principle, is to be rationally supported as the guiding principle for our social moral system. There seems to be no discrepancy between egoism and utilitarianism, or between self-love and benevolence, at this level of public morality.

To me, this double-edged proof of utilitarianism seems to be the most convincing argument of all the attempts to demonstrate the supremacy of utilitarian ethical theory. When I explain to myself or to someone else the reason why we have to consider the interests of others, I will

surely use two types of arguments, either to directly appeal to the benevolence that we naturally have, or to argue that we need to consider others in order to attain our own self-interest. However, as we will see in the following sections, 'the most convincing' argument does not mean that it is perfectly convincing.

11.3.1 Slight differences in social moral systems

In fact, Brandt states that there will be some differences between the 'utilitarian' social moral system which a person supports out of benevolence and the one which is endorsed from selfish motives (see Brandt 1979, pp. 207, 221). For example, our benevolent preferences would urge us to support a social moral system that requires everyone to consider the interests of others whether they will bring us any benefit or harm. However, our selfish preferences may prompt us to choose a social moral system that requires us to consider only the interests of the members of 'our' society or group, in which people have a reciprocal relationship. It would be fair to say that nowadays – more precisely, as far as current generations and those in the near future are concerned – almost all people on earth affect each other, and in this sense, we can see *some* reciprocal relationships between any two parties in this global community. It would be wise then, even from an egoistic viewpoint, to apply the same utilitarian social moral code even to people in regions and countries that we are not currently familiar with. For all that, once a class system is established in a society, for instance, there is always a possibility that upper-class people will disagree on whether to expand their utilitarian consideration to powerless lower-class people. There are other questions as to whether we should expand our utilitarian consideration even to nonhuman animals and to remote future generations. Our benevolent nature would encourage us to take them into moral consideration with no reservations, but our egoistic nature would try to convince us that it is pointless to consider their interests since they will never be able to benefit us. Thus the courses of actions prescribed by people who support a utilitarian *social moral system* while maintaining egoistic or utilitarian *personal ethical views* can significantly differ on some important issues of social concern.

The least we can say is that a person would always rationally support utilitarianism as the social moral code that should be prevalent in a society in which all the members have reciprocal relationships. Brandt appears to be content with this conclusion. However, especially today, when we have to deal with environmental issues that will most seriously affect people who will come 50 or 100 years after us, it makes

a big difference in whether we can convincingly claim that we ought to consider future generations. These potential differences between benevolence-based and egoism-based social moral systems are critical. We have to admit that some loose ends still remain.

I am not attempting to solve this problem in the present book. However, there is one thing that we can learn from Brandt's argument. In order to convince myself that I ought to consider the interests of future generations, I can take either of two paths: to appeal to my own benevolent nature by making full use of my imagination and compassion, *or* to make myself believe that considering their interests will somehow benefit me in the long run.

11.3.2 Internal conflicts still remaining

However, the possibility of disagreement on social moral codes is not the only problem we have as to the so-called dualism of practical reason. Even when we suppose that there are few differences between the forms of a utilitarian social moral system endorsed by benevolent and egoistic people, another problem remains unsolved.

The conclusion we reached in the previous section can be called a 'practical' harmony of egoism and utilitarianism at the level of public morality. This 'harmony' by no means implies that one is reducible to the other or that both turn out to be based on a common fundamental principle. It simply means that utilitarianism would be widely adopted as the guiding principle for a social moral code that we should act upon, whether our personal ethical views are egoistic or utilitarian.

However, Sidgwick's problem of dualism cannot be fully settled by this 'harmony'. What Sidgwick meant by the practical conflict between egoism and utilitarianism was that *in a person's mind* the conflict remains and continues to lead him astray every time he makes a practical decision as to what he ought to do in a particular situation. As we have already seen, a person would always support a utilitarian social moral system and show his (at least superficial) commitment to follow its code. However, this does not necessarily mean that he feels comfortable abiding by it, and a strongly selfish person would at times attempt to deviate from it. He may not allocate due consideration to people who are below him in social status and who are unlikely to fight against him. When he is confident that his misconduct will never be revealed, he may attempt to promote his own interests at the sacrifice of others. On the other hand, there is certainly a risk of enormous self-sacrifice in following a utilitarian social moral code. Thus, even though following the utilitarian social moral code and promoting self-interest may *generally*

be in harmony, they may not be in perfect accord on all occasions. This intrapersonal discrepancy was the very problem Sidgwick was most concerned about.

As for the possibility of such deviation from utilitarian morality, here are some points as to why it is not beneficial to deviate from it. First, it is never prudent to ignore other people's interests even though these people are lower in status, because the tables could be turned, and because, even if everyone's social status remains unchanged, it is always possible for others to take revenge if they become desperate. Second, your conviction that your misconduct will never be revealed very often proves illusory. Especially when your misconduct *harms* a person and leaves him feeling disgusted and furious, the victim will remember, as long as he is alive, the fact that he was harmed by someone and will do everything to identify the offender; thus your misbehavior is very often revealed sooner or later. Even if the fact that *you* performed the offending action remains unknown, people's recognition of the harm itself, which must have been done by *someone*, could significantly damage your self-interest. This is because we become even more anxious and suspicious of people around us when the offender remains unidentified. It is easy for mutual distrust and excessive vigilance to grow in such a situation, which will make society much less tranquil and productive even for the offender himself. Third, as for the claim that utilitarianism and egoism cannot coincide because the former could order a person to sacrifice himself for a greater social good, we can suggest that utilitarians would generally recommend us to avoid self-sacrifice even for the sake of a *seemingly* greater good. This is because, in reality, the great pain that the person must experience when sacrificing himself, plus the great sorrow and/or frustration that many compassionate people would have by observing his self-sacrifice, would significantly diminish the total happiness or satisfaction in most cases. Nevertheless, all these ideas only suggest a loose correlation of the observance of a social moral code with the pursuit of self-interest. They never guarantee a perfect harmony between them.

Brandt himself attempts to alleviate the conflict between utilitarian and egoistic ways of thinking by suggesting the following four points. First, it is – generally, and in the long run – beneficial for a person to maintain a disposition to observe a social moral code. Second, people normally have some degree of compassion or sympathy, and caring for others will bring satisfaction for such people. Thus considering the interests of others is in such benevolent people's self-interest. Third, an egoistic person could perhaps benefit from his immoral act if it were

never revealed, but it requires considerable wiliness to keep such an act from being detected. Fourth, cases in which a person really has to tragically sacrifice himself for the sake of others' interests are relatively rare (Brandt 1996, p. 290 f.). However, Sidgwick already noticed all these points. Being perfectly aware of a *general* correlation between self-love and benevolence, Sidgwick was concerned about the undeniable fact that the discrepancy between them remains.

Brandt himself admits that, 'In some cases, the traditional problem of conflict between self-interest and morality to some extent remains, and even the problem of what it is rational to do about such cases' (Brandt 1996, p. 302). After all, the puzzle of this dualism is not fully solved even today. This means, again, that contemporary utilitarians such as Brandt have not yet overcome the problems Sidgwick presented. The acute insights of Sidgwick, who recognized the real problem of the dualism of practical reason, are clearly demonstrated here as well.

However, in order to bring the problem of dualism closer to solution and to corroborate Sidgwick's statement that 'I do not mean that if we gave up the hope of attaining a practical solution of this fundamental contradiction, [. . .] it would become reasonable for us to abandon morality altogether' (*ME* p. 508), I will make the following modest yet constructive remark before finishing this chapter. First, as Brandt neatly demonstrated, we can meaningfully claim that a utilitarian social moral system would supersede the egoistic one. Second, in a prudent person, egoistic motives would urge him to deviate from utilitarian morality on very limited occasions. Thus, at least when we are dealing with issues of public concern, we can rationally expect that most people may agree to discuss them using utilitarian methods, or at least that they would show their benevolent consideration for others in their discussions. We can also expect that rational people will generally act according to a utilitarian social moral code once it is adopted, even though they may sometimes be tempted to deviate from it. The possibility of the conflict between self-love and benevolence or between egoism and utilitarianism remains, but these two modes of thinking are not as much on a par with each other as we initially expected.

Concluding Chapter

We have finished the examination of Sidgwick's ethical theory and its implications for contemporary utilitarianism. In this concluding chapter, let us simply summarize the arguments we developed in the previous chapters.

Utilitarian ethical theory is typically comprised of several unique factors, such as *consequentialism*, the *maximization* principle, *hedonism* and the policy to express the aggregation of pleasures as its *sum total*. We have elucidated how these components are analyzed and sustained by Sidgwick. According to him, ***consequentialism*** is supported through our critical examination of common-sense morality and our intuitive comprehension of the fundamental moral Principles of Rational Self-Love and Benevolence. That is, we concluded that consequentialism is the only viable way to systematically make a moral decision, by observing that nonconsequentialist approaches cannot help us to coherently determine the rightness and wrongness of an act, and by arguing that we ultimately appeal to various goods that an act will bring about when we consider what action one ought to take. The ***maximization*** principle of **the *sum total* of people's goods** is derived from an analysis of the concept of 'good' and the combination of two intuitive principles, Self-Love and Benevolence, *plus* the assumption that 'the whole' is to be construed as the sum total of its parts. ***The hedonistic interpretation of 'the ultimate good'*** is proved to be most plausible by first clarifying the concepts of pleasure and good and then examining all conceivable candidates for the ultimate good. This hedonistic value theory supports the idea that 'the good on the whole' should be understood as the sum total of pleasures, and thus the utilitarian principle of maximizing the sum total of people's pleasure is derived. Finally, the overall plausibility of utilitarian theory is confirmed by our well-considered commonsense.

By examining all these aspects of Sidgwick's argument, I have clarified the exact structure, content and foundations of his utilitarian theory as much as I can.

We also investigated Sidgwick's conceptual analyses, the four basic conditions for valid reasoning, and the three fundamental moral principles, all of which serve as the basis of ethics. One crucial point to be made regarding these analyses is Independent Interpretation, which shows the true significance of the essential distinction between the Principle of Justice and the other two fundamental moral principles that was previously underestimated.

Based on such explication of Sidgwick's ethics and his utilitarianism, we further inquired into contemporary discussions concerning (1) the fundamental principles of utilitarian ethical theory, (2) hedonistic versus preference-based value theories, (3) interpersonal comparisons of the strength of preferences, and (4) the possibility of reconciling egoism with utilitarianism. In doing this, we examined how contemporary utilitarians have developed new arguments to overcome the theoretical difficulties of utilitarianism and of ethics in general.

One feature that contemporary utilitarians inherited from Sidgwick is the resolution to develop an unbiased moral theory without appealing to moral intuitions and/or so-called commonsense that are often ambiguous, conflicting and dogmatic. In the same spirit Hare attempted to clarify the philosophical foundations of utilitarian ethical theory by using the minimum tools, such as language and facts, that we share. In his linguistic analysis, Hare spelled out that Sidgwick's Principle of Justice actually states the logical property (universalizability) of ought-judgments. He also demonstrated the link between prescriptivity of moral judgments and the concept of preferences. These points certainly contributed to the clarification of the structure of utilitarian ethical theory. However, Hare's utilitarian theory, which appeared to be based only on the logic of ought-judgments and recognized facts, turned out to be imperfect. In using the conditional reflection principle to represent other people's preferences to oneself, and in assuming that one ought to give equal weight to all the preferences represented to oneself, Hare's theory implicitly introduces additional requirements concerning the *quantitative* treatment of goods, which are equivalent to Sidgwick's Principles of Self-Love and Benevolence. It is important to remember that we could elucidate this flaw in Hare's theory by bringing to light the true significance of Sidgwick's three principles and the differences among them. Only because we reexamined Hare's argument with the knowledge of Sidgwick's ethics could we notice the unresolved issues

of contemporary utilitarianism and the challenges we should deal with in the future. Hare's moral philosophy contains a hurdle that has not yet been overcome. We should provide further justification for the fundamental principle(s) which cannot be explained simply as the logical requirement of ought-judgment. We need to prove, or at least explain, why one ought to give equal weight even to the preferences of others or to those at different times according to the strengths of those preferences.

Next we examined the preference-satisfaction theory that some contemporary utilitarians came to adopt in place of classical hedonism. Once again we focused attention on Hare's argument. It seemed that the preference-satisfaction theory is more favorable than hedonism in that it does not require the intricate 'proof' that we needed for hedonism, and yet it is practicable and easy to apply. When we scrutinized the recent Hajdin–Hare debate, however, it turned out that, at least for some preferences, the state of mind in which those preferences are either satisfied or frustrated is the key to moral reasoning. I claimed that this is another example which shows the singular quality of Sidgwick's ethical theory. One main point in Sidgwick's hedonism was that the state of mind (feeling) in which one's preference is satisfied is what we should ultimately consider in our moral reasoning. This is exactly what Hajdin claimed about Hare's preference theory, but Sidgwick further attempted to provide the *proof* of it, in order to explain the reason *why* such a state of mind is the key to moral reasoning. This is the argument that Hare did not provide. Sidgwick's proof of hedonism should be regarded as a laudable attempt to fill the gap between our moral motives and the representation of other people's preferences to ourselves.

As for the problems regarding the interpersonal comparison of the strength of preferences, we first discussed the difficulties that arise when we deny this kind of comparability, referring to arguments after Sidgwick. We then examined Harsanyi's claim that we need some process of converting each person's preference scale into a certain common scale if we are to make an accurate interpersonal comparison of the strength of preferences. It is Harsanyi's great contribution that he contrived a more detailed structure of interpersonal comparison than Sidgwick, who simply postulated that we can quantify the intensity of various people's pleasure. The device that Harsanyi proposed to measure and compare people's preferences at different times is the method of imagining the exchange of positions, representing another's preferences to oneself, and forming one's own extended preferences. Still, the theoretical difficulties of interpersonal comparison remain, for there is

no guarantee that a person's extended preferences or his own conversion ratio agree with those of others. Moreover, even if we assume that the strengths of preferences are interpersonally comparable, we can still question if the utilitarian principle is the only principle that can be advocated. Utilitarianism is the claim that we ought to give equal weight to people's preferences according to their strengths and to maximize the sum total of people's preference-satisfaction. There are other theorists, however, who claim different methods to weigh people's preferences, wishes, needs, life plans, etc. in a fair manner. Some attempt to consider people's preferences without making any of the interpersonal comparisons advocated by utilitarians. Others are inclined to adopt nonutilitarian ways of comparing people's goods, such as Rawls's difference principle and the principle of multiplying people's utilities. One of our future tasks would be to do some comparative study of these utilitarian and nonutilitarian principles of weighing and/or balancing people's different goods.

As for the well-known dualism of practical reason, or the discrepancy between utilitarianism and egoism, we examined Brandt's approach to solve this dilemma at the level of public morality. Whether or not our personal ethical views are egoistic, we will accept a utilitarian social moral code if we are fully rational in Brandt's sense. Thus we can establish this limited supremacy of utilitarianism in the field of public morality. Admitting all this, however, there remains the discrepancy between what my egoistic personal view tells me to do and what my utilitarian social code orders me to do, and I will continue to feel this internal conflict about what I really ought to do. Moreover, there is another problem, which is that a highly selfish person might not extend his 'utilitarian' social code to cover future generations, who are unable to reward him or revolt against him. So the dualism of practical reason is still alive today, as Sidgwick predicted.

Through our reexamination of contemporary utilitarianism, summarized above, we found that contemporary discussions have provided some better explanations of the problems Sidgwick left us with. More frequently, however, I showed that Sidgwick's argument is, contrary to our initial expectations, more sound and precise than that of contemporary thinkers. When we compare Sidgwick's ethical theory with modern-day ones, we notice that Sidgwick's argument is quite up-to-date in its basic framework and is based on even more acute analyses, which take various perspectives into consideration. Sidgwick had already developed the analyses which are essential components of contemporary utilitarianism (e.g., the analysis of moral judgments and the three fundamental

moral principles). In Sidgwick's argument we even find points that have been overlooked by recent thinkers (the need to introduce the Principles of Self-Love and Benevolence, types of preferences that are to be morally considered, etc.). Sidgwick presented detailed discussions on crucial theoretical difficulties that are still unresolved (the dualism of practical reason and the interpersonal comparison of preferences). We have not yet surpassed Sidgwick. Rather, only by investigating Sidgwick's arguments can we clarify the structure of contemporary utilitarian ethics, reveal its crucial difficulties and address future tasks facing contemporary moral philosophers.

Sidgwick's *Methods of Ethics* contains far more topics than I could cover in the present book – the issue of free will, the relationship of ethics to politics, criticism of evolutionary ethics, distributive justice, analyses of Kant, Green and other philosophers, etc. I believe, however, that even this limited examination of Sidgwick's ethical theory has shown the remarkable significance of his ideas for contemporary moral philosophy. The present book has elucidated the theoretical foundations, strengths and problems of utilitarian ethics as much as possible, through the explication of Sidgwick's theory and its relation to contemporary discussions. By doing so, I clearly emphasized the importance of assimilating the great ideas of our distinguished predecessor.

Notes

Introduction

1. As will be further explained in note 3 of Chapter 6, the idea of this interpretation was first advocated by Professor Emeritus Soshichi Uchii, and developed by Nakano-Okuno, who discussed this topic with him in 1997. In the Japanese version of this book, this interpretation appeared as 'Uchii-Okuno Interpretation', but was renamed as Independent Interpretation at the suggestion of Prof. Uchii in January 2011.
2. As far as I know, the only exception is Uchii's 1998 paper, in which he criticized Hare based on his interpretation of Sidgwick as stated in note 1.

1 The Scope of Ethics

1. The reason why Sidgwick did not make such a distinction is suggested in *ME* Book III, Chapter 7, Section 1, where he discusses the classification of common virtues and duties. There Sidgwick points out that classifying duties, as well as virtues, into social and self-regarding ones does not precisely reflect our common moral sense for the following reasons. First, the distinction between social and self-regarding duties seems to be drawn by considering whether the *consequences* of acts affect others or the agent himself, but at least some common moral rules apparently order certain acts without referring to any ulterior consequences. Second, almost all actions bring about various effects both to others and to the agent himself, and we have to select among relevant effects to make the above distinction. It is hard, however, for our commonsense to discern which effects are significant and which ones are not. Furthermore, some virtues, such as courage, can be exercised both for egoistic and social purposes. According to the classification stated above, we would have to discuss such virtues as both social and self-regarding ones, and that means dividing the same virtue into two categories. For this reason it is problematic to classify the rules of common virtues or duties into individual acts performed in public and those that are done in a purely personal domain. Thus Sidgwick, who wanted to start his discussion with the examination of commonsense morality, did not adopt this classification. However, in 11.2.1 of the present book I will argue that when we configure the proper scope of utilitarianism, it is helpful to use the distinction between a personal ethical view that an individual embraces in his mind and a social moral code that he must follow in public.
2. On page 5 of *ME* Sidgwick shows the idea that a rational act must be determined by referring to a certain principle. This requirement is not unnatural, because we turn to ethics to obtain a systematic guide for actions.

2 An Overview of *The Methods of Ethics*

1. In doing so, Sidgwick recognized that he must base his examination of the three methods and the fundamental moral principles on rigid analyses and verification, rather than dogmatically claiming the validity of a particular method. This stance is clearly demonstrated in Sidgwick's criticisms of his contemporary thinkers. For example, he could not agree with James Martineau, an intuitionist who attempted to solve moral issues by referring to a ranking table of various motives for actions. He did not sympathize with Thomas Hill Green's theory of self-realization either, pointing out that it is very ambiguous which actions would lead to the fulfillment of one's own potential. Sidgwick was also critical of evolutionary ethics, as advocated by Herbert Spencer and Leslie Stephen, which was very popular in his time; he strongly suspected that not everyone would agree with the ultimate grounds for their claims.
2. For example, Book IV, Chapter 3 of the first edition of *The Methods of Ethics* was titled 'The Proof of Utilitarianism (continued)', whereas the same chapter in the seventh edition is renamed 'The Relation of Utilitarianism to the Morality of Common Sense'. The former gives us the impression that the validity of utilitarianism can be proved when we examine our common-sense morality. Additionally, though the ending of *ME* is essentially the same in both editions in that Sidgwick confesses his inability to reconcile utilitarianism and egoism, the first edition presents only two fundamental moral principles, that is, the Principle of Justice and the Principle of Rational Benevolence (considered to be the foundation of utilitarianism), and does not contain the Principle of Rational Self-Love, which is stated in the seventh edition as the essential component of egoism. The readers of the first edition of *ME* might have had the impression that utilitarianism has a certain philosophical basis, in the form of fundamental principles, while egoism does not have such solid theoretical grounds.
3. Some may argue that Sidgwick criticized and eventually rejected dogmatic intuitionism. However, by examining the method of utilitarianism, Sidgwick came to reaffirm the importance of having certain moral rules for our daily life, assuming that those rules are similar to the existing rules of common-sense morality. Thus Sidgwick did not completely deny the main claim of dogmatic intuitionism, that is, the importance of observing ordinary rules.
4. Schneewind and Peter Singer correctly understand this point. See Schneewind 1977, Ch. 6 pp. 191–2 and Singer 1974. Donagan sees *ME* as a somewhat inconsistent work which attempted to defend utilitarianism but failed to establish its validity. This opinion may perhaps sound close to Sidgwick's own voice, but it fails to grasp the intent of *The Methods of Ethics*, which confined itself to the impartial criticism and explication of all three methods. Donagan in Schultz 1992, p. 447.
5. Schneewind regards the relationship between morality and Christianity as one of the main subjects of *The Methods of Ethics* (Schneewind 1974, p. 391). Even if this was true, however, I am not going to read *ME* from a religious perspective, for I strongly believe that Sidgwick's argument on the essential features of ethics can be fully understood and accepted by people from all walks of life, with or without religious beliefs.

3 Three Methods, Intuition, and Commonsense

1. R. M. Hare addresses this point. According to him, 'universal' is the opposite of 'singular', and 'general' is the opposite of 'specific' (*MT* 2.6. See also 8.2.1 of the present book). However, it seems to me that in our daily use of language, the term 'general' does not necessarily mean 'nonspecific', but very often simply means 'widely accepted' or 'prevalent'.
2. As for the reason why he listed happiness, perfection/excellence and virtues (duties) as the ultimate reasons for action, Sidgwick leaves us with the comment that 'This threefold difference in the conception of the ultimate reason for conduct corresponds to what seem the most fundamental distinctions that we apply to human existence' (*ME* Bk. 1 Ch. 6 p. 78). These distinctions include the one between *the conscious being* and *the stream of conscious experience*, the latter of which can further be divided into *action* and *feeling*. According to Sidgwick, excellence or perfection is the ideal object of human development, in which a human individual is regarded as a continuous being. Duties or virtues denote the actions that ought to be done, and thereby reflect the fact that humans have various experiences and perform different kinds of action. Happiness means a set of desirable feelings, and thus assumes that humans have feelings (*ME* Bk. 1 Ch. 6 Sec. 1 p. 78. It is said that Sidgwick added this part for the second edition. See Schneewind 1977, Ch. 6 p. 199). By this Sidgwick probably means that the classification of the ultimate reasons into happiness, perfection and virtues (duties) holds true because it corresponds to these distinctions about human existence. However, this explanation does not sound convincing to me. A more important argument for identifying the widely accepted ultimate reasons for conduct would be the one in which Sidgwick examines and eliminates other candidates for such ultimate reasons besides the four described above.
3. Schneewind insists that Sidgwick actually envisioned the fourth method, that of pursuing perfection, and that there was no need to include it in the method of intuitionism (Schneewind 1977, p. 204). For Sidgwick, however, it was hard to identify the method of aiming for perfection independently of other methods. In a footnote of Book I, Chapter 2 of *ME*, Sidgwick states as follows: 'I omit, for the present, the consideration of the method which takes Perfection as an ultimate end: since, as has been before observed, it is hardly possible to discuss this satisfactorily, in relation to the present question, until it has been somewhat more clearly distinguished from the ordinary Intuitional Method' (*ME* p. 20 fn.). It is not a serious defect that he did not discuss the method of perfection independently. What Sidgwick addressed in *ME* were mainly two things, that is, the reconciliation of utilitarianism with common-sense morality, and the discrepancy between utilitarianism and egoism. For his purpose, it was probably sufficient to categorize the conformity with virtues/duties as versions of common-sense views and discuss them in Book II of *ME*, and to provide an additional examination of the concept of perfection in his discussion of the ultimate good and the proof of hedonism.
4. Sidgwick does not clearly define the term 'principle' in *ME*. However, he seems to be using this term in the sense we explained here, judging from his use of 'principle(s)' in *ME* Bk. 1 Ch. 1 pp. 5–6 and 8, in the title of Bk. 1 Ch. 6,

and in other places where Sidgwick talks about the principle and the method of egoism or utilitarianism.

5. Some theorists suggest that Sidgwick's terminology for these three methods is somewhat confusing. While egoistic and universalistic hedonism indicate the feature that makes an act the right one (i.e., pleasure as the ultimate end of action), intuitionism suggests the feature that is displayed in the act of making a moral judgment (wherein one makes a judgment without further inferences). Thus it is claimed that the two types of hedonism should be paired with deontology and that intuitionism should be coupled with 'Inferentialism' or something of the sort (Raphael 1974). However, considering the fact that one of the greatest debates in Sidgwick's time was the one between intuitionists and utilitarians, his contemporaries would have better understood Sidgwick's terms.
6. M. G. Singer maintains that two methods can come into conflict (or contradict each other) only when their designated ends are one and the same (M. G. Singer 1974, p. 441). We might use the term 'conflict/contradict' in this way, but I believe this does not grasp what Sidgwick intended to convey.

4 Meta-Ethical Analyses

1. It is interesting that Sidgwick already noticed this difficulty in the first edition of *ME* and stated as follows: 'In discussing whether moral distinctions are perceived by the Reason, it is especially important to make clear the point at issue. As we know nothing of any faculty of the mind except from its effects, and only assume different faculties to explain or express differences among the mental phenomena which we refer to them, we must always be prepared to state what characteristics in the feeling or cognition investigated such reference imports' (*ME*1 pp. 22–3). This passage also suggests the idea that the faculty of reason is what is inferred from our actual judgments, acts and mental phenomena, or what is assumed in order to explain these activities. Sidgwick seems to have revised this passage in the seventh edition simply to avoid misunderstanding.
2. The term 'ethical' stated here must include, besides morality in a narrow sense, the use of 'ought' which denotes prudence in egoism. This is because Sidgwick regards the dictate of self-love as a manifest duty and includes it within the scope of ethics. See 1.1 and 3.1 of the present book, and *ME* p. 386. As we will see, however, in our later analysis we will focus on the use of 'ought' in a narrow moral sense, in which it is easier to see the logical properties peculiar to 'ought' judgments.
3. To say this does not mean that we can *define* 'ought' or 'right' as equivalent to good or the greatest good. The concept of the greatest good implies, by definition, what ought to be aimed for (if at all attainable). However, the reverse, that the concepts of 'ought' or 'right' are by definition what will bring about the greatest good, cannot always hold true. For example, according to dogmatic intuitionism, the right thing to do is unconditionally dictated regardless of the good it will bring about. In this regard, Schneewind is wrong in stating that 'We need therefore only say that the right act is [. . .] the best act which it is possible for the agent to do, to make clear the way in which the

same basic notion is involved in both concepts [i.e., "right" and "good"]' and 'rightness is defined in terms of bringing about the greatest good within the agent's power' (Schneewind 1977, pp. 225 and 307). Sidgwick clearly asserts that the concepts of 'ought' and 'right' are indefinable. There remains a gap between the concept of 'right' and that of the greatest attainable good. For example, while the former contains an explicit dictate, the latter only arouses a mild desire. Thus rightness cannot be officially defined by using the term good. However, as Sidgwick says, it is possible to clarify the *relationship* of the notions of 'right' or 'ought' to other notions (*ME* p. 33). Here we have admitted that the proposition stated in this section is analytically true by noticing a certain relationship between two concepts, that is, that 'the greatest good' implies that it 'ought' to be sought.

4. My analysis differs from that of Shionoya or Schneewind. Shionoya seems to simply presuppose the maximization principle of good (Shionoya 1984, p. 171), but in my opinion the maximization principle is not what everyone takes for granted nor what can be easily derived. Schneewind claims that by examining common-sense morality we notice that the maximization of good is what makes the right act right (Schneewind 1977, p. 308), but this argument is not convincing either. Through his examination of common duties and virtues, Sidgwick certainly admits that an act is approved as right only when it is related to *some* good, but this is not equivalent to the claim that the *maximization* of good is right. There is a possible objection to my view, such as that of Bernays, whom Schneewind cites. I will not detail this objection, but it roughly claims that Sidgwick would not endorse such a claim since he always avoids tautological truths (see Schneewind 1977, p. 308). I believe I fully responded to this objection in 4.3.2 and 5.2 of the present book.
5. The original text says 'to mean what I *should* practically desire if my desires were in harmony with reason, assuming my own existence alone to be considered' (Emphasis added). It would be worthwhile to point out that the term 'should' used here does not have the mandatory connotation of 'ought'. This is because, first, the dictate 'ought' is not supposed to manifestly appear in the definition of good, and, second, the potential mandate of the notion of good is already expressed by the phrase 'if my desires were in harmony with reason'. Schneewind and Christiano refer to the same passage and rephrase it as 'to mean what I *whould* practically desire' (Schneewind 1977, p. 224; Christiano in Schultz 1992, p. 264. Emphasis added). I thank Tetsuji Iseda for suggesting this point to me.
6. Schneewind makes this point in Schneewind 1977, p. 369, claiming that it clearly shows that Sidgwick regards the good on the whole as the aggregation of the goods of individuals. However, Schneewind does not show how this aggregation principle is logically derived from the idea explained so far. We will elucidate the structure of this derivation in the present book.

5 Testing the Significance of Apparent Truths

1. Sidgwick's 1879 paper, 'The Establishment of Ethical First Principles' (Sidgwick 1879a), explains the reason why he seems to have abruptly presented these four conditions in *ME*. In this paper Sidgwick states that, if we are to establish

the first principles of ethics, the premise of our reasoning must contain a certain norm, which is expressed by the term 'ought'; according to him, we cannot determine ethical first principles that should be applied to everyone just by exploring facts. Then he goes on to claim that there are only two methods for us to attain such fundamental principles. One is to start off with a defined proposition that we clearly recognize as true and then to eliminate arbitrary limitations to arrive at a simpler and more comprehensive proposition. To take Sidgwick's own example, we can start by admitting the truth of the proposition, 'all suffering of rational human beings should be avoided', and then, upon noticing that it is arbitrary and groundless to cling to the question of whether or not an individual in pain is rational, we may come to acknowledge a simpler proposition, 'all suffering should be avoided'. The other method is to first establish a set of general criteria for discerning true principles from false ones, and then to explore fundamental principles while referring to these criteria. Obviously, Sidgwick followed the second procedure in presenting these four conditions.

2. According to Sidgwick 1879a, a self-evident proposition is one which we can appropriately recognize without reference to other propositions.
3. See Schneewind 1977, Ch. 2, p. 64. In Sidgwick 1879a, along with discussing the condition of Cartesian clarity and distinctness, Sidgwick examines what Reid and others proposed as the conditions of universal (or almost universal) acceptance and primitiveness (i.e., being based on a primitive belief). Sidgwick himself is critical of the condition of primitiveness.
4. However, a doubt may arise that the Principles of Rational Self-Love and Benevolence, which Sidgwick presents as self-evident principles, do not satisfy this condition. We will discuss this point later.

6 The Three Fundamental Principles

1. Thus the Principle of Justice simply requires us to impartially apply the logic of ought judgment to everyone; this is quite different from so-called distributive justice, which requires us to allocate equal portions of good to all people.
2. For Sidgwick, the quantitative comparison of various goods will be expressed, *via* his proof of hedonism, by the comparison of the greatness of pleasures. In doing so, he simply assumes that pleasures are measurable and their greatness (quantity) is mutually comparable. The task of solving the problem of measuring and comparing the amounts of goods for different individuals at different points in time will be taken over by John Harsanyi. We will deal with Harsanyi's argument in 10.2 of this book.
3. I owe this claim to Professor Soshichi Uchii. I have to admit that I did not understand the point of his criticism against a very common interpretation that the Principles of Justice, Self-Love and Benevolence are simply three modes of applying exactly the same principle, a topic which Uchii addressed in group email communications among Japanese co-translators of *The Methods of Ethics* (September 6–8, 1997). After he repeatedly emphasized this point, I gradually came to understand the differences among the three principles. Then I finally realized the crucial significance of these differences

when I reexamined Hare's argument from Sidgwick's perspective. Thus I am completely indebted to Professor Uchii for the basic idea of Independence Interpretation, though I believe I made my own contribution to the development and promotion of this interpretation by verifying it through in-depth analysis of the whole text of *ME*, and by clearly illustrating its importance in the context of reevaluating contemporary utilitarianism. Uchii's own argument on this topic, together with his own criticism of Hare's universalizability, is fully developed in Uchii 1998.

7 Philosophical Foundations of Utilitarianism

1. One possible interpretation is that these assumptions are finally upheld when Sidgwick concludes, via his proof of hedonism, that the ultimate good amounts to 'pleasure that each individual feels at each point in time' and nothing else. If pleasure is the *sole* ultimate good, 'the ultimate good on the whole' would have to be the simple aggregation of pleasures; there would be no such thing as an 'extra' good that makes 'the good on the whole' more than the aggregation of its parts, for allowing such extra goods would be to introduce nonhedonistic values into argument. Sidgwick's proof of hedonism will be discussed in 7.2, and the derivation of the utilitarian principle of maximizing the sum total of pleasures will be examined in 7.3 of this book.
2. Desire usually means an impulse that surges when pleasure is yet to exist. Here, however, 'desire' is used as the most suitable term to express a felt stimulus to one's will, as just described. It would also be useful to draw the readers' attention to the differences between a desire/motive and a will, which I explained in 1.2.2 of this book. A will makes a conscious choice among desires or motives and determines one single action to be taken. An individual can have multiple desires at once, and each of those desires stimulates one's will to make a choice among them.
3. For instance, my desire that 'the world's carbon dioxide emissions in 2020 be reduced by 25 percent compared to those in 1990' is being *fulfilled* if the desired state of affairs occurs in 2020 (that is, if the CO_2 emissions in that year is actually 25 percent less than 30 years ago). This holds true even if I am ignorant of the fact of whether it really happened and therefore do not have any changes in my feelings.
4. In fact, Sidgwick states that it may also be difficult to compare two pleasures felt simultaneously. A person sometimes experiences two or more types of pleasures at the same time, as when he listens to music while drinking wine. In such cases, however, we are often unable to adequately weigh those two pleasures. This is partly because, in such cases, the sources of those pleasures interfere with each other so that both pleasures do not exhibit their normal intensity; in most cases a person is concentrating either on the taste of the wine or on the sound of the music, and when he experiences both, his sensibility tends to lessen. More frequently, the two pleasures mix together and are felt by a person as one single pleasant state of consciousness, so that he cannot evaluate them separately (*ME* p. 141). In the present book, however, I will focus on the difficulties of comparing the pleasures of various people

at different times, and will not discuss the comparison of pleasures that an individual simultaneously feels at a certain time.

5. A similar classification is already presented and explained by Yuichi Shionoya, though his terminology and his use of it are different from mine (Shionoya 1984, pp. 168–70). I adopted this classification here because I think this is certainly indispensable in order for us to understand Sidgwick's hedonism. I listed four main subcategories (i–iv) in the main text, but, theoretically, there could be others, such as those that consider pleasures of a limited or an indefinite number of people.
6. Sidgwick especially distinguishes between psychological egoistic hedonism and ethical universalistic hedonism because he discovered these two elements in J. S. Mill's utilitarianism.
7. Sidgwick's 1873 article, 'John Stuart Mill', suggests that Sidgwick highly evaluated J. S. Mill in many respects, but not in regards to Mill's ethical theory. One reason for this presumably lies in Sidgwick's recognition that Mill based his proof of ethical universalistic hedonism on psychological egoistic hedonism. However, one could question whether J. S. Mill really advocated psychological egoistic hedonism as Sidgwick understood it. Mill certainly insists that one always pursues one's own pleasure (again, the phrase 'one's own pleasure' here should be understood as one's satisfied feelings in a very broad sense), but he does not seem to claim that one always *consciously* pursues one's own pleasure. Sidgwick mainly denies this latter claim of *conscious* pursuit of one's own pleasure in his criticism of psychological hedonism. In contrast, Mill fully admits that a person sometimes pursues the attainment of virtues or even self-sacrifice, having little or no awareness that he is pursuing his own pleasure. See Mill 1863. I thank Makoto Suzuki for giving me many suggestions as to the interpretation of J. S. Mill through our discussions in 1997–8.
8. Even if we admit that we may often attain pleasure by not consciously pursuing it, it is still meaningful to insist that we 'ought to' pursue pleasure and to actually strive for it. The paradox of hedonism simply teaches us that, when we follow ethical hedonism and attempt to attain pleasure, it is sometimes advisable to avoid directly aiming for it.
9. According to Schneewind, Book III, Chapter 14 of *ME* became its final version in the fifth edition after numerous alterations. Those changes mainly involve integrating arguments from other chapters, expanding on certain points, and rearranging the arguments within the chapter. Schneewind thinks, however, that the central arguments and the main conclusion of Chapter 14 are quite consistent throughout all seven editions of *ME*. Thus I will chiefly use the seventh edition in the present book.
10. From the arguments discussed above, we can speculate how Sidgwick would have replied to the problem of a so-called pleasure machine. It is often said that hedonism is wrong because we would surely regard this machine as deeply undesirable and unacceptable even if it produces a great amount of pleasure (see Smart and Williams 1973, p. 19 ff). A pleasure machine is an imaginary machine which continuously provides a person with extremely pleasant stimuli via electrodes stuck into his head. (We do not need such an eccentric device if we just imagine a drug that continues to produce quite strong pleasures for the rest of one's life.) These pleasures are supposed to

be very intense, and can be provided repeatedly and constantly. Critics of hedonism claim that hedonists must admit that a life connected to this machine is highly desirable, while most ordinary people do not think so. However, Sidgwick would counterargue this criticism as follows. What we evaluate as 'undesirable' in this example is not the *feeling* which the person on such a machine is experiencing. When we examine this feeling apart from all external relationships, the feeling itself is certainly desirable – that is why we still call it pleasure. What we judge to be undesirable is the 'objective condition' in which this feeling is produced, that is, the relationship between the feeling and the machine causing it. Sidgwick would further argue that, upon reflection, we would soon notice that we cannot justify our judgment that this objective relationship is undesirable unless we refer to still other types of pleasant or unpleasant feelings. For example, some may claim that the future pleasures which this person could have experienced were he not connected to that artificial device might have been far more colorful, diversified and pleasurable. Others may suggest that he would have felt disgusted upon learning that he was being manipulated by the machine, or that people who watched him and realized that a person could be so manipulated would have felt great displeasure. These pleasures and displeasures should be weighed separately from the feelings of the person on the machine in order to determine the overall desirability of using such a pleasure machine.

11. The phrase 'unconsciously utilitarian' implies that we do not always consciously exercise utilitarian thinking in our daily life. However, we are certainly conducting a utilitarian thought process when we attempt to solve conflicts or justify the general validity of particular duties by explicitly or implicitly balancing people's overall pleasure and pain – even if we do not outrightly identify ourselves as utilitarians.

Part II A Reexamination of Contemporary Utilitarianism

1. The point that the dualism of practical reason might undermine the self-evidence of the Principles of Self-Love and Benevolence was first suggested in Seth 1901, p. 180. The same point is discussed in M. G. Singer 1974, p. 446. However, these authors seem to believe that the discrepancy between egoism and utilitarianism suggests a clear-cut conflict between the Principles of Self-Love and Benevolence. Indeed, Sidgwick himself, in discussing this dualism, states that 'a harmony between the maxim of Prudence and the maxim of Rational Benevolence must be somehow demonstrated' (*ME* p. 498); he thereby gives the impression that the former maxim is exclusively for egoism and only the latter is for utilitarianism. However, as we analyzed them, the Principle of Self-Love is essentially the principle to ignore *the difference in time* in our treatment of pleasures, and the Principle of Benevolence is the principle to ignore *the difference of particular individuals* in our treatment of pleasures (at least this is how I interpreted Sidgwick's theory). According to utilitarianism, we are supposed to ignore not only the differences of individuals but also those of time; thus utilitarianism must presuppose both principles of Self-Love and Benevolence as we interpret them.

8 An Approach not Appealing to Moral Intuition

1. Although Rawls would not admit this, Hare regards Rawls's theory as appealing to moral intuitions. Hare then criticizes such 'crypt-intuitionists', claiming that 'If one goes through such writings and discounts all the arguments which rest on undefended moral intuitions of substance, nothing is left but the mere moral opinions of the authors with which they hope we will agree'. See *MT* 4.4, p. 76.
2. Hare discusses almost nothing about his interpretation of Sidgwick in *Moral Thinking*. In its preface he simply states that 'In particular, I am, when I speak of intuition, neither attacking nor defending Sidgwick; my intended targets are more recent, and will be recognized' (*MT* p. vi). We can gather two things from this brief statement, however. Hare was not defending Sidgwick probably because Hare, unlike Sidgwick, thought it proper *not* to rely on *any* moral intuitions in developing his own argument. We may recall that Sidgwick's 'philosophical intuitions' are not *linguistic* intuitions on which Hare bases his own analysis but *moral* intuitions that have substantial contents. Still, Hare was never attacking Sidgwick because Hare derives theses identical to Sidgwick's from Hare's own linguistic analysis.
3. To Hare, 'universal' means the attribute of not referring to any particular individual, and it does not necessarily mean 'not specific' or 'general'. A judgment with specific details can be a universal one if it contains no reference to proper nouns or particular individuals. For example, the simple judgment, 'do not kill anyone', and the very specific judgment, 'do not kill anyone *except* in self-defense and judicial executions', are both universal in Hare's sense, in that they do not refer to particular individuals.
4. Skeptics often doubt the existence of other minds and their preferences. Hare, however, asserts that this topic is not peculiar to utilitarian ethics but a profound problem for philosophy in general, and proposes to set it aside while admitting that it is an unresolved issue. He then shows sympathy toward the analogical argument that 'we reasonably guess that beings so like us in all other respects are also like us in having similar conscious experiences under similar conditions' (*MT* 7.2), and proceeds on the assumption that other people's preferences do exist. Sidgwick makes the same assumption, and I will do the same.
5. Despite this measure of Hare's, the problem of the interpersonal comparison of preferences is still often addressed. Griffin and others, for example, continue to claim that what one can learn about another's preferences from observation is, at best, the order of his preferences (namely, which he prefers over which) and not the intensity of his preferences (how strongly he prefers something). See Seanor and Fotion 1988, pp. 73–88; Iseda 1996, p. 28. Kenneth Arrow holds the same point. If their claims are correct, it is surely impossible for a person to compare the strength of his own preference with that of someone else. This issue is discussed in Chapter 10 of the present book.
6. These two difficulties have already been suggested by several researchers, and I was especially inspired by a paper written by Tetsuji Iseda (Iseda 1996). I agree with him on many points, but not on all. Iseda seems to interpret R. M. Hare as if Hare believed that in making a universalizable judgment one

must be able to accept the judgment in question *at the moment of imagining oneself being put in the position of someone else*. Iseda criticizes Hare on this account, but in my opinion that is not what Hare claimed in the first place.

9 A Reappraisal of Hedonism

1. Hedonism and the preference-satisfaction theory correspond to what R. B. Brandt called Happiness Theory and Desire Theory respectively, or what Derek Parfit called Hedonistic Theory and Desire-Fulfillment Theory respectively (Brandt 1979; Parfit 1986). Smart presents a similar classification about utilitarianism (Smart 1978). Preference-based utilitarianism is advocated by R. M. Hare, Jan Narveson and many scholars in the field of economics. Of course, there are some who still support the hedonistic version of utilitarian ethics. One such hedonistic utilitarian is Brandt. I will mention Brandt's view again in note 3 of this chapter.
2. The term 'external preferences' was coined by Dworkin, and Hare renamed it as 'non-experiential preferences' in his later article (Hare 1998). In the same article, Hare also used the term 'asynchronic preferences' for what he used to call *now-for-then* preferences. In the following, I will use 'external preferences' and 'asynchronic preferences' to denote the two kinds of preferences simply because I prefer concise words.
3. R. B. Brandt's support for hedonism is based on a simple guess. According to him, the observation of our common-sense behavior reveals that what we wish to do for the sake of others is not to cause a state of affairs that satisfies their preferences but to increase their happiness (Brandt 1979, p. 148). More precisely, he presents further grounds, which are based on a psychological theory of sympathy; but that is not perfectly convincing. Interestingly, in Chapter 5 of *A Theory of the Good and the Right*, Brandt wavers to proclaim that happiness for oneself and others is the *only* thing that we regard as ultimately desirable. (Here 'ultimately desirable' should be construed to be what we would desire and pursue *per se* if we had maximum knowledge as to facts and logic.) Brandt believes that we will, upon careful reflection, continue to regard our own pleasures and other people's pleasures as desirable, considering the basic make-up of our minds, especially the mechanism of our sympathy. However, Brandt admits that we will continue to regard certain other things as desirable as well. Then it is unclear why Brandt can clearly support hedonism, while ignoring other desirable items.

10 Interpersonal Comparison and Maximization

1. Strictly, the ultimate target of these preferences would be *a certain feature* of the state of affairs resulting from an act, and in most cases, such a feature would be pleasure or happiness. However, I will use the phrase 'preference *for the state of affairs* that would result from an act'. This is because preferences are relevant to our moral judgments in that they affect our choice among *the states of affairs* we could bring about (or our choice among acts that would bring about those states of affairs). Thus my preference for the *pleasure* that I will experience when a fight is over will be expressed as my preference for

the state of affairs in which a fight is over – or more simply, my preference for ending the fight. In making a moral judgment, this preference would urge me to choose an act that is most likely to end the fight. Though we cannot strictly identify 'a preference for a feature which is involved in a state of affairs' with 'a preference for that state of affairs', there is a close relationship between these two, similar to a means–end relationship, and we would be allowed to assume that if we have the former type of preference then we will also have the latter preference. However, when we lose sight of the ultimate target of our preferences, which is not actually the whole state of affairs but *some feature* that is involved in that state of affairs, we have to undergo the process of reconfirming what the ultimate good is for us, which was the main topic of the preceding chapter.

2. Strictly speaking, it is assumed here that a person's preferences are always mutually consistent in following this ranking. If person P has the preference of X > Y > Z, then P must always prefer X over Y, Y over Z and X over Z. This means that each individual must satisfy the principle of transitivity as to his own preferences. Some of my critics have insisted that human preferences are often irrational and not transitive – they claim that they could prefer, for example, the movie *Spider-Man One* over *Spider-Man Two*, Two over Three but Three over One. (I doubt that such a claim simply means they rank all three equally or capriciously.) If a person's preference were completely random, however, it would be impossible and futile to consider people's preferences in making any kind of judgment. It would be far more productive to assume that people's preferences are, at least ideally, consistent. The same is true of the consistency of moral judgments. If my 'moral' preferences (which will be explained in the following) do not satisfy the rule of transitivity, my very attempt to make a moral judgment becomes futile.
3. I found this point suggested in Shibata 1988, pp. 266–7; Saeki 1980, p. 121. However, Saeki mentions this in the context of introducing the point made by another welfare economist Yew-Kwang Ng. Saeki himself is critical of considering the strengths of preferences.
4. This move is also useful for me to compare the preferences of two other persons, or to compare two preferences of the same individual who is not myself. However, imagining such hypothetical situations would be even more difficult, and therefore I would check the correctness of my perception by referring to other people's preference scales that can be obtained by the Neumann-Morgenstern method. For more details, see Harsanyi's original argument.

Bibliography

Listed below are literature on Sidgwick's ethics and contemporary utilitarianism. I attempted to make a sufficiently detailed list of Sidgwick's works on ethics and moral philosophy, while often omitting his articles on economics, education, literature, etc. To obtain a more comprehensive list that includes all the articles and manuscripts written by Sidgwick, see bibliographies in Schneewind 1977 and Sidgwick 2000 (edited by Marcus G. Singer).

Those works with asterisk (*) are the ones I frequently referred to in developing my arguments in the present book. I also included several books that appeared after the Japanese version of my book was published in 1999. Though these books had no direct influence on my arguments in the present volume, I believe they should be included here considering their academic importance.

1. Works by Sidgwick

* Sidgwick, Henry (1907) *The Methods of Ethics*, 7th edition (London: Macmillan); 1st edition, 1874. The 7th edition was reprinted in 1962 (Chicago: the University of Chicago Press), and in 1981 with a foreword by John Rawls (Indianapolis: Hackett). Japanese translation of the 5th edition by R. Nakajima, T. Yamabe and H. Ohta in 1898 (Tokyo: Dai-Nippon-Tosho); German translation of the 7th edition by C. Bauer in 1909 (Leipzig: Klinkhardt); Italian by M. Mori in 1995 (Milano: Il Saggiatore).

—— (1866) 'Ecce Homo', *Westminster Review*, Jul. 1866, 58–88. Reprinted in Sidgwick 1904.

—— (1871a) Review of John Grote's *Examination of the Utilitarian Philosophy, Cambridge University Reporter*, 8 Feb. 1871,182–3.

—— (1871b) Review of John Grote's *Examination of the Utilitarian Philosophy, The Academy*, Apr. 1871, 197–8.

—— (1871c) 'Verification of Beliefs', *Contemporary Review*, Jul. 1871, 582–90.

—— (1872a) 'Pleasure and Desire', *Contemporary Review*, Apr. 1872, 662–72.

—— (1872b) 'The Sophists' I and II, *The Journal of Philology*, Vol. 4–5, No. 8–9, 1872–3. Reprinted in Sidgwick 1905.

—— (1873) 'John Stuart Mill', *The Academy*, May 1873, 193.

—— (1874) 'On a Passage in Plato's *Republic*', *The Journal of Philology*, Vol. 5, No. 10, 274–6.

—— (1876a) 'The Theory of Evolution in its Application to Practice', *Mind*, Vol. 1, No. 1, 52–67.

—— (1876b) 'Philosophy at Cambridge', *Mind*, Vol. 1, No. 2, 235–46.

—— (1876c) 'Review of F. H. Bradley's *Ethical Studies*, *Mind*, Vol. 1, No. 4, 545–9.

—— (1877a) 'Hedonism and Ultimate Good', *Mind*, Vol. 2, No. 5, 27–38.

—— (1877b) 'Bentham and Benthamism in Politics and Ethics', *Fortnightly Review*, N. S. Vol. 21, 627–52. Reprinted in Sidgwick 1904.

—— (1879a) 'The Establishment of Ethical First Principles', *Mind*, Vol. 4, No. 13, 106–11.
—— (1879b) 'The So-Called Idealism of Kant', *Mind*, Vol. 4, No. 15, 408–10.
—— (1880a) 'Kant's Refutation of Idealism', *Mind*, Vol. 5, No. 17, 111– 14.
—— (1880b) Review of Fouillee's *L'Idée moderne du droit*, *Mind*, Vol. 5, No. 18, 135–9.
—— (1880c) 'Mr Spencer's Ethical System', *Mind*, Vol. 5, No. 18, 216–26.
—— (1882a) 'On the Fundamental Doctrines of Descartes', *Mind*, Vol. 7, No. 27, 435–40.
—— (1882b) 'Incoherence of Empirical Philosophy', *Mind*, Vol. 7, No. 28, 533–43. Reprinted in Sidgwick 1905.
—— (1882c) Review of L. Stephen, *The Science of Ethics*, *Mind*, Vol. 7, No. 28, 572–86.
—— (1883a) 'A Criticism of the Critical Philosophy I', *Mind*, Vol. 8, No. 29, 69–91.
—— (1883b) 'A Criticism of the Critical Philosophy II', *Mind*, Vol. 8, No. 31, 313–37.
—— (1883c) 'Kant's View of Mathematical Premises and Reasonings', *Mind*, Vol. 8, No. 31–2, 421–4 and 577–8.
—— (1883d) *The Principles of Political Economy* (London: Macmillan). 3rd edition, 1901.
—— (1884) 'Green's Ethics', *Mind*, Vol. 9, No. 34, 169–87.
—— (1885a) Review of Fowler's Progressive Morality, *Mind*, Vol. 10, No. 38, 266–71.
—— (1885b) Review of J. Martineau's *Types of Ethical Theory*, *Mind*, Vol. 10, No. 39, 426–42.
—— (1886a) 'Dr. Martineau's Defence of *Types of Ethical Theory*', *Mind*, Vol. 11, No. 41, 142–6.
—— (1886b) 'The Historical Method', *Mind*, Vol. 11, No. 42, 203–19.
—— (1886c) 'Economic Socialism', *Contemporary Review*, Vol. 50, 620–31. Reprinted in Sidgwick 1904.
—— (1886d) *Outlines of the History of Ethics* (London: Macmillan). 5th edition, 1902. Reprinted by Thoemmes Press, 1993.
—— (1887) 'Idiopsychological Ethics', *Mind*, Vol. 12, No. 45, 31–44.
—— (1888) 'The Kantian Conception of Free Will', *Mind*, Vol. 13, No. 51, 405–12.
—— (1889) 'Some Fundamental Ethical Controversies', *Mind*, Vol. 14, No. 56, 473–87.
—— (1890) 'The Morality of Strife', *International Journal of Ethics*, Vol. 1, 1–15. Reprinted in Sidgwick 1898 and Sidgwick 1919.
—— (1891) *The Elements of Politics* (London: Macmillan). 3rd edition,1908; 4th edition, 1919.
—— (1892a) 'The Feeling-Tone of Desire and Aversion', *Mind*, N. S. Vol. 1, No. 1, 94–101.
—— (1892b) Review of H. Spencer's *Justice: Being Part VI of the Principles of Ethics*, *Mind*, N. S. Vol. 1, No. 1, 107–18.
—— (1893a) 'My Station and its Duties', *International Journal of Ethics*, Vol. 4, No.1, 1–17.
—— (1893b) 'Unreasonable Action', *Mind*, N. S. Vol. 2, No. 6, 174–87. Reprinted in Sidgwick 1898.

—— (1894a) 'Luxury', *International Journal of Ethics*, Vol. 5, No. 1, 1–16.
—— (1894b) 'A Dialogue on Time and Common Sense', *Mind*, N. S. Vol. 3, No. 12, 441–8. Reprinted in Sidgwick 1905.
—— (1895a) 'The Philosophy of Common Sense', *Mind*, N. S. Vol. 4, No. 14, 145–58. Reprinted in Sidgwick 1905.
—— (1895b) 'Theory and Practice', *Mind*, N. S. Vol. 4, No. 15, 370–5.
—— (1895c) Review of D. G. Ritchie's *Natural Rights*: *A Criticism of Some Political and Ethical Conceptions*, *Mind*, N. S. Vol. 4, No.15, 384–8.
—— (1896) 'The Ethics of Religious Conformity', *International Journal of Ethics*, Vol. 6, No. 3, 273–90.
—— (1898) *Practical Ethics* (London: Swan Sonnenschein).
—— (1899) 'The Relations of Ethics to Sociology', *International Journal of Ethics*, Vol. 10, No. 1, 1–21. Reprinted in Sidgwick 1904.
—— (1900) 'Criteria of Truth and Error', *Mind*, N. S., Vol. 9, No. 36, 8–25. Reprinted in Sidgwick 1905.
—— (1901) 'The Philosophy of T. H. Green', *Mind*, N. S., Vol. 10, No. 1, 18–29.
—— (1902) *Philosophy: Its Scope and Relations, An Introductory Course of Lectures*, with an editorial note by James Ward (London: Macmillan).
—— (1902) *Lectures on the Ethics of T. H. Green, Mr. Herbert Spencer, and J. Martineau*, with a preface by E. E. Constance Jones (London: Macmillan).
—— (1903) *The Development of European Polity*, edited by E. M. Sidgwick (London: Macmillan).
—— (1904) *Miscellaneous Essays and Addresses*, edited by E. M. Sidgwick and A. Sidgwick (London: Macmillan).
—— (1905) *Lectures on the Philosophy of Kant and Other Philosophical Lectures and Essays*, edited by James Ward (London: Macmillan).
—— (1919) *National and International Right and Wrong: Two Essays*, with a preface by James Bryce (London: George Allen & Unwin). This is the reprint of 'Public Morality' and 'The Morality of Strife' from Sidgwick 1898.
—— (2000) *Essays on Ethics and Method*, edited, with an introduction, by M. G. Singer (Oxford: Oxford University Press).

Complete works

Sidgwick, Henry (1996a) *The Works of Henry Sidgwick*, with a new introduction by John Slater (Bristol: Thoemmes Press). 15 volumes.
—— (1996b) *The Complete Works and Select Correspondence of Henry Sidgwick* (Past Masters Series), edited by Bart Schultz (Charlottesville, VA: InteLex Corporation).

2. Works on Sidgwick

Albee, Ernest (1901) *A History of English Utilitarianism* (New York: Macmillan).
Bain, Alexander (1876) 'Mr. Sidgwick's Methods of Ethics', *Mind*, Vol. 1, No. 2, 179–97.
Blanshard, Brand (1974) 'Sidgwick the Man', *The Monist*, Vol. 58, No. 3, 349–70.
—— (1984) *Four Reasonable Men* (Middletown, CT: Wesleyan University Press).
Bradley, F. H. (1877) 'Mr. Sidgwick on 'Ethical Studies',' *Mind*, Vol. 2, No. 5, 122–6.

—— (1877) *Mr. Sidgwick's Hedonism: An Examination of the Main Argument of The Methods of Ethics* (London: Henry S. King &Co.).
Broad, C. D. (1930) *Five Types of Ethical Theory* (London: Kegan Paul). Reprinted in 2000, 2001 (London: Routledge).
Bryce, James B. (1903) *Studies in Contemporary Biography* (London: Macmillan).
Bucolo, P., Crisp, R. and Schultz B. (eds) (2007), *Henry Sidgwick: Happiness and Religion – Proceedings of the World Congress, University of Catania* (Catania: Universita degli Studi di Catania).
Calderwood, H. (1876) ' Mr. Sidgwick on Intuitionalism', *Mind*, Vol. 1, No. 2, 197–206.
Christiano, Thomas (1992) 'Sidgwick's Desire, Pleasure, and Good', in Schulz 1992, pp. 261–78.
Darwall, Stephen L. (1974) 'Pleasure as Ultimate Good in Sidgwick's Ethics', *The Monist*, Vol. 58, No. 3, 475–89.
Deacon, Richard (1985) *The Cambridge Apostles* (London: Robert Royce Ltd).
Donagan, Alan (1977) 'Sidgwick and the Whewellian Intuitionism: Some Enigmas', *Canadian Journal of Philosophy*, Vol. 7, No. 3, 447–65. Reprinted in Schultz 1992.
Ezorsky, Gertrude (1974) 'Unconscious Utilitarianism', *The Monist*, Vol. 58, No. 3, 468–74.
Frankena, William (1974) 'Sidgwick and the Dualism of Practical Reason', *The Monist*, Vol. 58, No. 3, 449–67.
Geninet, H. (2009) *Politiques Comparees: Henry Sidgwick et la politique moderne dans les 'Eléments Politiques'* (Reims: Université de Reims Champagne-Ardenne). In French.
Gizychi, G. von (1891) Review of *The Methods of Ethics* by H. Sidgwick, *International Journal of Ethics*, Vol. 1, 120–1.
Green, T. H. (1877) 'Hedonism and Ultimate Good', *Mind*, Vol. 2, No. 6, 266–9.
Hayward, F. H. (1900) 'The True Significance of Sidgwick's Ethics', *International Journal of Ethics*, Vol. 11, No. 2, 175–87.
—— (1901) *The Ethical Philosophy of Sidgwick* (London: Swan Sonnenschein & Co.). Reprinted in 1993 (Bristol: Thoemmes Press).
James, D. G. (1970) *Henry Sidgwick: Science and Faith in Victorian England* (London: Oxford University Press).
Mackenzie, J. S. (1891) Review of *The Methods of Ethics* by H. Sidgwick, *International Journal of Ethics*, Vol. 1, 512–4.
Nakajima, Rikizo (1908) *Sidgwick's Ethical Theory* (Tokyo: Dobunkan). In Japanese.
Nakano-Okuno, M. (1998a) 'Rethinking Sidgwick's Hedonism', *Arche*, Kansai Society of Philosophy, No. 6, 38–48. In Japanese.
—— (1999) 'Butler and Sidgwick on Intuitionism', in Shigeru Yukiyasu (ed.), *Modern British Ethics and Religion: Butler and Sidgwick* (Kyoto: Koyo Shobo), pp. 204–23. In Japanese.
—— (2002) 'An Essay concerning Freedom of the Will, Morality, and Responsibility: The Contrast between Kant and Sidgwick', *Annals in Philosophy*, vol. 61 (Fukuoka, Japan: Dept. of Philosophy, Faculty of Humanities, Kyushu University), 75–93. In Japanese.
—— (2007) 'Sidgwick and Kant: On the So-Called "Discrepancies" between Utilitarian and Kantian Ethics', in P. Bucolo, R. Crisp and B. Schultz (eds) 2007, pp. 259–333.

Raphael, D. D. (1974) 'Sidgwick on Intuitionism', *The Monist*, Vol. 58, No. 3, 405–19.
Rashdall, Hastings (1885) 'Professor Sidgwick's Utilitarianism', *Mind*, Vol. 10, No. 38, 200–26.
Schneewind, J. B. (1974) 'Sidgwick and the Cambridge Moralists', *The Monist*, Vol. 58, No. 3, 371–404.
* —— (1977) *Sidgwick's Ethics and Victorian Moral Philosophy* (Oxford: Oxford University Press).
* Schultz, Bart (ed.) (1992) *Essays on Henry Sidgwick* (Cambridge: Cambridge University Press).
—— (2004) *Eye of the Universe: An Intellectual Biography* (Cambridge: Cambridge University Press).
Seth, J. (1901) 'The Ethical System of Henry Sidgwick', *Mind*, N. S., Vol. 10, No. 1, 172–87.
Shimamoto, Ainosuke (1919) 'Green versus Sidgwick: On the Logical Meaning of Ethics', *Essays in Honor of the 25th Anniversary of Professor Nakajima's Professorship* (Tokyo: Meguro-Shoten). In Japanese.
Shionoya, Yuichi (1983) 'The Structure of Sidgwick's Utilitarianism', *Hitotsubashi University Research Series, Economics*, Vol. 24, 117–214. In Japanese.
* —— (1984) *The Structure of Values in Economic Philosophy: Utility versus Rights (Kachi Rinen no Kozo)* (Tokyo: Toyo Keizai Inc). In Japanese.
Sidgwick, Arthur and Eleanor Mildred (1906) *Henry Sidgwick: A Memoir* (London: Macmillan).
Singer, Marcus George (1974) 'The Many Methods of Sidgwick's Ethics', *The Monist*, Vol. 58, No. 3, 420–48.
Singer, Peter (1974) 'Sidgwick and Reflective Equilibrium', *The Monist*, Vol. 58, No. 3, 490–517.
Sorley, W. R. (1901) 'Henry Sidgwick', *International Journal of Ethics*, Vol. 11, No. 2, 168–74.
Stephen, Leslie (1901) 'Henry Sidgwick', *Mind*, N. S., Vol. 10, No. 1, 1–17.
Tsunashima, Ryosen (1922) 'Sidgwick's Methods of Ethics, Book I', *Complete Works of Ryosen Tsunashima*, Vol. II (Tokyo: Shunjusha). In Japanese.
Uchii, Soshichi (1988) *The Law of Freedom, The Logic of Interests* (Kyoto: Minerva). In Japanese.
—— (1998) 'Sidgwick's Three Principles and Hare's Universalizability', http://www1.kcn.ne.jp/~h-uchii/sidg&hare_index.html. Also in Uchii (1999) 'Three Essays on Ethics', *Memoir of the Faculty of Letters*, Vol. 38, 118–46.
Williams, Bernard (1982) 'The Point of View of the Universe: Sidgwick and the Ambitions of Ethics', *Cambridge Review*, 7 May 1982, 183–91. Reprinted in B. Williams (2006) *The Sense of the Past: Essays in the History of Philosophy* (Princeton: Princeton University Press).
Wodehouse, Helen (1907) 'The Idealist and the Intuitionist', *International Journal of Ethics*, Vol. 17, No. 2, 164–80.
* Yukiyasu, Shigeru (ed.) (1992) *Studies on H. Sidgwick* (Tokyo: Ibunsha). In Japanese.
Yukiyasu, Shigeru (1975) 'The Life and Ethics of H. Sidgwick', *The Bulletin of the Okayama College of Science*, Vol. 10, 41–52. In Japanese.
—— (1990) 'J. Rawls' Critique of Sidgwick's Theory of Justice, and its Development', *Bulletin of Faculty of Education, Okayama University*, Vol. 83, 143–54. In Japanese.

—— (1994) 'The Criticism of Utilitarianism and the Acceptance of Self-Realization in Modern Japan: Japanese Traditional Ideas in Ryosen Tsunashima and Kitaro Nishida', *Bulletin of Faculty of Education, Okayama University*, Vol. 97, 103–20. In Japanese.

3. Works on contemporary utilitarianism and others

Arrow, K. J. (1951) *Social Choice and Individual Values* (New York: John Wiley). 2nd edition, 1963. Reprinted in 1970 (New Haven and London: Yale University Press).

Ayer, A. J. (1954) 'The Principle of Utility' in his *Philosophical Essays* (London: Macmillan). Partly Reprinted in Glover 1990, pp. 48–51.

Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A. and Tipton, S. M. (1996) *Habits of the Heart: Individualism and Commitment in American Life* (Berkeley: University of California Press).

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

—— (1992) *Morality, Utilitarianism, and Rights* (Cambridge: Cambridge University Press).

—— (1996) *Facts, Values, and Morality* (Cambridge: Cambridge University Press).

Cumberland, Richard (1672) *De Legibus Naturae*. Selections with parallel translation in D. D. Raphal (ed.) (1969) *British Moralists 1650–1800*, Vol. I (Oxford: Clarendon Press), pp. 77–102.

Eguchi, Satoshi (1994) 'Hare's Utilitarianism and External Preferences', *Studies on Ethics* (*Rinrigaku Kenkyu*), The Kansai Society for Ethics, Vol. 24, 107–19. In Japanese.

Frankena, William (1988) 'Hare on the Levels of Moral Thinking' in Seanor and Fotion 1988, pp. 43–56.

Fukaya, S. and Terasaki, S. (eds) (1983) *The Nature and Aspects of the Good* (Kyoto: Showado). In Japanese.

Glover, Jonathan (ed.) (1990) *Utilitarianism and Its Critics* (New York: Macmillan).

* Hajdin, Mane (1990) 'External and Now-for Then Preferences in Hare's Theory', *Dialogue* 29, 305–10.

Hare, R. M. (1961) *The Language of Morals* (Oxford: Oxford University Press).

—— (1963) *Freedom and Reason* (Oxford: Oxford University Press).

—— (1976) 'Ethical Theory and Utilitarianism' in H. D. Lewis (ed.), *Contemporary British Philosophy* (London: George Allen & Unwin). Reprinted in Hare 1989, pp. 212–30.

* —— (1981) *Moral Thinking* (Oxford: Oxford University Press).

—— (1989) *Essays in Ethical Theory* (Oxford: Oxford University Press).

—— (1991) 'Universal Prescriptivism' in Singer 1991, pp. 451–63.

—— (1994) 'Applied Philosophy and Moral Theory: R. M. Hare Talks to Philosophy Today (Interview)', *Philosophy Today* 38, No. 17, 3–6.

—— (1998) 'Preferences of Possible People' in C. Fehige and U. Wessels (eds), *Preferences* (Berlin: de Gruyter), pp. 399–405.

Harsanyi, John C. (1955) 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economy* 63, 309–21. Reprinted in Harsanyi 1976.

—— (1976) *Essays on Ethics, Social Behaviour, and Scientific Explanation* (Dordrecht: Reidel). Reprinted in 1980.

—— (1977) *Rational Behaviour and Bargaining Equilibrium in Games and Social Sciences* (Cambridge: Cambridge University Press).

Hastings, A. W. (1980) 'Utilitarianism' in J. Hastings (ed.) (1980) *Encyclopedia of Religion & Ethics* (Edinburgh: T & T Clark), pp. 558–67.

Hirano, Hitohiko (1985) 'Natural Rights Theories and Utilitarianism: Concerning R. M. Hare's Moral Theory', *Ho Tetsugaku Nenpo*, Japan Association for Legal Philosophy, 61–77.

Hobbes, Thomas (1651) *Leviathan*. Reprinted in 1985 (London: Penguin Classics).

Iseda, Tetsuji (1996) 'Preference Utilitarianism and Universalizability Thesis', *Bulletin of the Japan Society of British Philosophy*, Vol. 19, 21–34. In Japanese with an English abstract.

Ito, Kunitake (1997) *The Philosophy of Human Rationality* (Tokyo: Keiso-shobo). In Japanese.

Jeffrey, Richard C. (1965) *The Logic of Decision* (Chicago: University of Chicago Press). 2nd edition, 1983.

Kamino, Keiichiro (1969) 'The Concept of Utility in Ethical Theories', *The Journal of Philosophical Studies (The Tetsugaku Kenkyu)*, the Kyoto Philosophical Society, Vol. 510, 25–43.

Kato, Hisatake (1997) *Introduction to Contemporary Ethics* (Tokyo: Kodansha). In Japanese.

—— (2009) *Ethics of Consensus Formation* (Tokyo: Maruzen). In Japanese.

Kawamoto, Takashi (1995) *The Adventures of Contemporary Ethics (Gendai Rinrigaku no Boken)* (Tokyo: Sobunsha). In Japanese.

Kupperman, Joel J. (1982) 'Utilitarianism Today', *Revue internationale de philosophie* 141, 318–30.

Lyons, David (1965) *The Forms and Limits of Utilitarianism* (Oxford: Oxford University Press).

—— (1992) 'Utilitarianism' in L. C. Becker (ed.) (1992) *Encyclopedia of Ethics* (New York: Garland), pp. 1261–8.

Lyons, William (1972) 'Is Hare's Prescriptivism Morally Neutral?' *Ethics* 82, No. 3, 259–61.

MacIntyre, Alasdair (1967) *A Short History of Ethics* (London: Routledge & Kegan Paul).

Mill, John Stuart (1863) *Utilitarianism*.

Moore, G. E. (1903) *Principia Ethica* (Cambridge: Cambridge University Press). Revised edition in 1993.

Nakano-Okuno, Mariko (1997 and 1998b), 'Parfit's Defense of Utilitarianism: A New Look Based on the Theory of Persons and Personal Identity', *The Journal of Philosophical Studies (The Tetsugaku Kenkyu)*, the Kyoto Philosophical Society, Vol. 564, 81–114 and Vol. 565, 84–100. In Japanese.

—— (1998c) 'Rationality of Action and Desire: An In-Depth Discussion of R. B. Brant's theory of Good', *Studies on Ethics (Kansai Rinrigaku Kenkyu)*, vol. 28, Kansai Society for Ethics, 110–22. In Japanese.

Narita, K. (1987) 'On the Universalizability of Moral Judgments', *Tetsugaku*, Keio University's Mita Philosophical Society, Vol. 84, 27–50. In Japanese.
Nozick, Robert (1974) *Anarchy, State, and Utopia* (New York: Basic Books).
Parfit, Derek (1984, 1986) *Reasons and Persons* (Oxford: Oxford University Press).
Rawls, John (1971) *A Theory of Justice* (Cambridge, MA: Harvard University Press).
Riker, William Harrison (1982) *Liberalism against Populism: A Confrontation between the Theory of Democracy and the Theory of Social Choice* (San Francisco: W. H. Freeman).
Saeki, Yutaka (1980) *The Logic of Social Decision-Making (Kimekata no Ronri)* (Tokyo; Tokyo University Press). In Japanese.
Sandel, Michael J. (2009) *Justice: What's the Right Thing to Do?* (New York: Farrar, Straus and Giroux).
Seanor, D. and Fotion, N. (eds) (1988) *Hare and Critics* (New York: Oxford University Press).
Sen, Amartya K. (1977) 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory', *Philosophy and Public Affairs* 6, 317–44.
Sen, Amartya and Williams, Bernard (eds) (1982) *Utilitarianism and Beyond* (Cambridge: Cambridge University Press).
Shibata, H. and Shibata, A. (1988) *Public Economics (Kokyo Keizaigaku)* (Tokyo: Toyo Keizai Inc.). In Japanese.
Singer, Peter (ed.) (1991) *A Companion to Ethics* (Oxford: Blackwell).
Smart, J. J. C. (1978) 'Hedonistic and Ideal Utilitarianism', *Midwest Studies in Philosophy*, Vol. 3, 240–51.
Smart, J. J. C. and Williams, B. A. O. (1973) *Utilitarianism: For and Against* (Cambridge: Cambridge University Press).
Sone, Yasunori (1982) *Contemporary Political Theories (Gendai no Seiji Riron)* (Tokyo: Chikumashobo). In Japanese.
Sprigge, T. L. S. (1988) *The Rational Foundation of Ethics* (London: Routledge & Kegan Paul).
Sugihara, S., Yamashita, S. and Koizumi, A. (eds) (1992) *Studies on J. S. Mill* (Tokyo: Ochanomizu Shobo). In Japanese.
Suzumura, Kotaro (1983) *Rational Choice, Collective Decisions, and Social Welfare* (Cambridge: Cambridge University Press).
Uchii, Soshichi (1974) 'Moral Reasoning', *Zinbun*, Vol. 13, 61–81.
—— (1974) 'On the Universalizability of Moral Judgements', *The Zinbun Gakuho*, Institute for Humanistic Studies, Kyoto University, Vol. 38, 19–39. In Japanese.
—— (1996) *Ethics and the Theory of Evolution* (Kyoto: Sekaishisho-sha). In Japanese. Its pdf version with English abstract is available at Uchii's website, http://web.me.com/uchii/Site/Papers.html.
Yamauchi, Tomosaburo (1991) *Putting Oneself in Another's Shoes: Hare's Moral Philosophy (Aite no Tachiba ni Tatsu)* (Tokyo: Keiso Shobo). In Japanese.
Yukiyasu, Shigeru (1988) 'R. M. Hare versus Defenders of Naturalistic Ethics', *Bulletin of School of Education, Okayama University*, Vol. 79, 13–23. In Japanese.

Index