ALTRUISM, CONFLICT, AND COOPERATION:

A STUDY OF TWO MIXED-MOTIVE MODELS OF STRATEGIC INTERACTION

By

AHMED JAMAL ANWAR

M.A. (Dalhousie)

A dissertation submitted for the degree of

DOCTOR OF PHILOSOPHY

(Department of Philosophy, Faculty of Arts

Registration No. 51/2015-16)

at the

UNIVERSITY OF DHAKA

July 2020

ALTRUISM, CONFLICT, AND COOPERATION:

A STUDY OF TWO MIXED-MOTIVE MODELS OF STRATEGIC INTERACTION

By

AHMED JAMAL ANWAR
M.A. (Dalhousie)

A dissertation submitted for the degree of

DOCTOR OF PHILOSOPHY

(Department of Philosophy
Faculty of Arts
Registration No. 51/2015-16)

at the

UNIVERSITY OF DHAKA

July 2020

Signature of Author and Date.....

DECLARATION

I declare that the thesis "*Altruism, Conflict, and Cooperation: A Study of Two Mixed-Motive Models of Strategic Interaction*" has been composed entirely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree or professional qualification at this or any other institution.

Name of Candidate: Ahmed Jamal Anwar, Reg. No: 51/2015-2016

Signature: Date:

DEDICATED

To the memory of my mother Anwara Begum and father Tasadduq Ahmad

ACKNOWLEDGEMENTS

First I would like to thank my supervisor, Professor Dr. Anisuzzaman, for his kind support, inspiration, generosity, patience, and above all his knowledge and skill as a philosophical generalist. I would also like to record my appreciation and thanks to Dr. M.A. Kashem who suggested me to register for the Doctoral program as a Faculty member of the University of Dhaka.

I would also like to thank my colleaguesDr. Md. Sajahan Mia, Dr. A.K.M. Haroon-ar Rashid, Dr. ShahKawthar Mustafa Abululayee , and Dr. M. MatiurRahman for their passionate concern and encouragement.

Special thanks to my colleagues and friends especially Mr. S.M. HumayunKabir, a former student of Jahangirnagar University and now serving as a Govt. college teacher of BCS cadre, who helped me in various ways in meeting the administrative functions and deadlines. I would also like to thank Mr. PanuGopal Pal, Sr. Administrative Officer of the Department of Philosophy, for his excellent cooperation in administrative matters.

Finally, I owe a great debt to my wife ShahanaParveen and my daughters Tahmina Anwar (Lecturer in English, Centennial College,Toronto), Dr. Tarana Anwar, and Farhana Anwar (Doctoral candidate, USA) for their constant support and inspiration.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF SYMBOLS

CERTIFICATE FOR COMPLETION OF DOCTORAL DISSERTATION

I certify that I have meticulously read the dissertation “Altruism, Conflict, and Cooperation: A Study of Two Mixed-Motive Models of Strategic Interaction” written entirely by Ahmed Jamal Anwar (Reg. No: 51/2015-2016) under my supervision, and that, in my opinion, it is fully adequate in scope and quality as a doctoral dissertation. I, therefore, recommend that the dissertation be placed to the examiners for evaluation.

Name of Supervisor: Prof. Dr. Anisuzzaman

C/o. Department of Philosophy

University of Dhaka

Signature: ……

Date: ……

Altruism, Conflict, and Cooperation: A Study of Two Mixed-motive Models of Strategic Interaction

CHAPTER ONE

INTRODUCTION

1.1 Preview

The purpose of this chapter is to introduce the present study which fully examines the main research problem formulated as a compound question as follows: Is altruism only a necessary condition, or else only a sufficient condition, or else either a necessary or a sufficient condition, or else both a necessary and a sufficient condition, or else neither a necessary nor a sufficient condition for mutual cooperation among rational

individuals interacting in circumstances where their interests are neither entirely identical nor completely contradictory, but rather mixed that leave open the possibility of cooperation as well as defection? We will argue that the right answer to the above question is that altruism is neither a necessary nor a sufficient condition for mutual cooperation to happen under the stated circumstances. But then we are naturally led to a second but nonetheless very important query which is as follows: If altruism is neither a necessary nor a sufficient condition, can it still be a 'contributory cause' of cooperation? Section 1.2 of this introductory chapter presents a brief description of the general background of the research problem. Next, Section 1.3 precisely defines the research problem, discusses the rationale for its investigation, and shows how the dissertation contributes to its solution. Section 1.4 which is the last section of this chapter presents a brief outline of the dissertation structure.

1.2 Background

Conflict and cooperation are inevitable interactions between two or more social units who engage one another to achieve their respective goals. These two types of interaction are diametrically opposed and yet so ubiquitous in social life that they may be regarded as the two opposite sides of the same coin. *Cooperation* is a sort of social interaction that happens when an individual or group incurs a cost to itself to help another individual or group obtain some personal benefit or attain a shared goal. *Conflict*, on the other hand, arises when individuals or groups work against each other and try to defeat an

opponent for achieving a greater share of private benefits[1]. The *reason* for the coexistence and ubiquity of conflict and cooperation as two different phenomena of society has been succinctly stated by John Rawls as follows:

> … although a society is a cooperative venture for mutual advantage, it is typically marked by a conflict as well as an identity of interests. There is an identity of interests since social cooperation makes possible a better life for all than any would have if each were to try to live solely by his own efforts. There is a conflict of interests since men are not indifferent as to how the greater benefits produced by their collaboration are distributed, for in order to pursue their ends they each prefer a larger to a lesser share. (1971, p. 126)

Thus it is obvious that both cooperation and conflict are common phenomena of collective life. Moreover, there is a profound significance of cooperation in collective life particularly on matters of mutual concern for a good purpose. Nevertheless, there are different or even conflicting views on the nature, extent, origin, and development of cooperation as well as its relationship with some associated concepts. Frequently, the distinctions between *cooperation* and *mutualism* and between *cooperation* and *altruism* as types of pro-social behavior and as opposed to antisocial behavior are not sufficiently made clear in the literature. As a result, these terms are often confused and misused. One of the confusions is that *altruism* and *cooperation* are supposed to be the *same* and not different types of behavior (e.g., Sussman & Cloninger, 2011). A second confusion concerns a popular and intuitively appealing but erroneous view supported by some eminent academics (Ball, 1985; Hamlin, 1986; Lea, Tarpy, & Webley, 1987; Samuelson

[1] It is noteworthy that the kind of conflict mentioned here is *interpersonal* conflict and is different from the *intra-personal* conflict that arises within a single individual because of the coexistence of incompatible attitudes towards objects in an environment. Intra-personal conflict is discussed in chapter 2.

& Nordhaus, 2006) who hold that in situations of *social dilemma* where there is a conflict between personal and public interests, *egoism*, i.e., the exclusive concern with one's self-interests, is the cause of conflict or noncooperation, and hence *altruism*, i.e., the absence of egoism, is the *cause* of cooperation implying that only unselfish behavior can create a cooperative culture and resolve the dilemma. A third confusion lies in confining altruism and cooperation to the lowest level of pro-social behavior by sticking to the inherent restrictions of the *biological definitions* instead of using the *psychological definitions* of the terms (Batson, 2011, 2014; Okasha, 2013).

To help advance knowledge and research in the area, this dissertation addresses the above confusions and errors and attempts to refute the wrong view concerning the causal connection between altruism and cooperation by formulating two different arguments to show that altruism is neither a necessary condition nor a sufficient condition for creating cooperation, despite the fact that it may sometimes serve as a contributing and critical condition for cooperation.

1.3 Problem Statement and Rationale for the Research

An essential characteristic of society is persistent interaction that refers to how different individuals or groups in the society act with or react to each other under different circumstances and thereby mutually influence each other's behavior. *Social interaction*, a central concept of sociology, has been defined by Macionis (2000) as "the process by which people act and react in relation to others" (p. 85). Sociologists have

described several basic patterns of social interaction[2] of which two in particular – cooperation and conflict – have received a great deal of attention in game theory which is the study of interdependent decisions of goal-oriented agents[3]. A number of distinguished thinkers (Hobbes, 1651; Malthus, 1798; Marx and Engels, 1848; Darwin, 1859; Keddy, 2001) presented a gloomy view of life characterized mainly by unconstrained selfishness, backstabbing, cheating, deception, duplicity, and constant conflict between individuals or groups of individuals over the access to limited resources. Society, however, cannot exist without some degree of cooperation in various spheres of life. Although the great evolutionary scientist Darwin (1859) made competition the main focus of his thinking, he himself was *puzzled* by and tried to draw our attention to the co-presence of cooperation that necessitates sacrifice and thereby reduces the fitness of the cooperator, on the one hand, and constant conflict and struggle for existence entailed by his theory of natural selection, on the other. In trying to resolve this Darwinian puzzle about the apparent incompatibility between selfishness and cooperation, a number of outstanding scientists or philosophers both before and after Darwin (Hobbes, 1651; Locke, 1689; Rousseau, 1762; Wynne-Edwards, 1962; Ardrey, 1970; Hamilton, 1964; Trivers, 1971; Rawls, 1971; Gauthier, 1986; Kohn, 1992; Tomasello, 2009; Wilson, 2014; Batson, 2011, 2014) offered various explanations of the emergence, continuity, and evolution of cooperation out of conflict in collective life.

[2] The other important and fundamental types of social interaction are exchange, competition, coercion, conformity, and accommodation (including compromise, truce, mediation, and arbitration).

[3] By an agent is meant anything such as a person, organization, machine, or software that decides on a means to achieve a given end.

Though animal and particularly human societies are replete with examples of highly cooperative behavior, the logic of its existence is difficult to understand. The source of this difficulty lies in our attempt at resolving the apparent incompatibility between individual rationality and collective rationality, or, between conflict and cooperation. As the famous prisoner's dilemma game shows, it often turns out that mutual cooperation is better than mutual defection for all the parties concerned, but unilateral defection, i.e., defection against cooperation, is even better than mutual cooperation for the defector while it yields the worst possible outcome for the unilateral cooperator, because non-cooperators tend to take advantage of cooperators by free-riding on the latter's cooperation, i.e., without bearing any cost in the form of reciprocal cooperation (Tucker,1950; Rapoport & Chammah, 1965). Consequently, all the parties who are assumed to be rational agents seeking to maximize their own interests would defect, and hence would have to settle for the outcome arising out of mutual defection which for all the parties involved turns out to be inferior to not only what was expected but also the moderate outcome that could be attained through mutual cooperation.

It is, however, evident from our everyday experiences that cooperation is a fundamental feature of social life. This has been concisely stated by Zaggl (2014) as follows: "Cooperation is the glue that binds individuals together and allows for the emergence of social structures on higher levels, such as families, groups, organizations, nations, and civilizations" (p. 197). And as scientists (e.g., Trivers, 1971; Axelrod & Hamilton, 1981; Kohn, 1992) point out, cooperation among members of the same or even different species is not only a possibility but a pervasive phenomenon.

The above situation might obviously lead to a pertinent question: Why do people cooperate? Cooperative behaviors, according to Butler (1729), are actually motivated by altruism, i.e., one individual's unselfish concern for promoting the welfare of another, while according to Hobbes (1651), Adam Smith (1759), and Becker (1974) they arise out of egoism, i.e., one individual's selfish and strategic concern for using cooperation merely as a means to ultimately promoting his or her own advantage or wellbeing? Thus in answering the question about the origin of cooperation both Butler (1729) and Hobbes (1651) accept the existence of cooperation, but while the former holds that it arises out of people's pro-social, i.e., unselfish, behavior toward others, the latter holds that it arises out of people's pro-self, i.e., selfish, behavior along with their interdependence, and rationality. Thus it is important to resolve the Hobbes-Butler debate on egoism vs. altruism which we attempt in chapter 2.

A more specific but intractable question, however, arises: *Why* do those who are presumed to be prudent people planning to persistently pursue personal preferences usually by acting against one another really indulge in cooperative or helpful behavior toward others even when they know that such behavior may be costly or harmful to them? As mentioned before and will be discussed in detail in Chapter 2, a one-shot play of the prisoner's dilemma would logically lead both the players to consciously settle for a deficient outcome through mutual defection while a mutually advantageous and cooperative outcome was available. One popular but intuitively appealing view that has also been endorsed by some respectable scholars, such as Ball (1985), Hamlin (1986), Lea, Tarpy, and Webley (1987), and Samuelson and Nordhaus (2006), attributes this regrettable failure to achieve the cooperative outcome to the egoistic choices of all the

individuals involved and holds that cooperation could be made possible by a pro-social type of behavior known as altruism. Thus they identify *altruism* as the cause of which *cooperation* is the effect.

Though the two terms "cooperation" and "altruism" have some resemblance in meaning and are often used concurrently, whether cooperation and altruism are *causally connected* or not remains an open question. The presumed causal connection between altruism and cooperation may, in terms of the distinction between necessary and sufficient condition, be interpreted in any of five different senses: (i) Altruism is a *necessary condition* for cooperation. (ii) Altruism is a *sufficient condition* for cooperation. (iii) Altruism is a necessary ***and*** sufficient condition for cooperation. (iv) Altruism is a necessary ***or*** sufficient condition for cooperation. (v) Altruism is *neither* a necessary ***nor*** a sufficient condition for cooperation.[4]

To clarify the differences among the five possible interpretations, let us analyze the meanings of the third, fourth, and fifth cases which are different ways of compounding the necessary and sufficient conditions. The third interpretation being a *conjunction* of the first and the second types is the strongest claim and the fourth one being a *disjunction* of the first and the second types is the weakest claim. The third interpretation holds that altruism is a *necessary* as well as a *sufficient* condition for cooperation and so implies that cause and effect are so related that they can be inferred from one another. This is indeed a common but rather strong claim. Scholars often confuse between *altruism* and *cooperation* and write as if these two terms are

[4] See (Copi, Cohen, & McMahon, 2014)) for a brief but illuminating discussion of the interpretation and use of the term "cause" in several different senses.

synonymous and so refer to the same type of behavior, though they are by definition different. There is no harm in this view as such except that the belief in this strong connection may lead some people to confuse between altruism and cooperation. For an example of the confusion, let us take a look at what Sussman and Cloninger (2011, p. 2) writes: "The concept of cooperation or of altruism (i.e., disinterested concern for another's welfare) is often assumed to be one of humanity's essential and defining characteristics." Thus the third possible explanation may potentially lead one to treat the two separate concepts of "altruism" and "cooperation" as synonymous. As a matter of fact, while altruism is essentially based on a concern for another's welfare, cooperation need not necessarily be based on a concern for the other and may even be based on a desire to promote one's own wellbeing. The third interpretation treats cause as a sufficient condition that includes all the necessary conditions which are extremely difficult or impossible to identify. Moreover, there may be alternative sets of sufficient condition for the same event where some parts of a sufficient condition may not be necessary. Hence, this interpretation is questionable.

The fourth interpretation being a *disjunction* of the first and the second types holds that altruism is a *necessary* or a *sufficient* condition for cooperation. This construal makes indeed a rather weak claim which is logically entailed by each of the three previous ones but implies none of them[5] and, unlike the third one is not based on any unsustainable assumption. The fifth interpretation is the negation of the fourth one and holds that altruism is neither necessary nor sufficient for cooperation. We will argue against the fourth possibility. But logically, this is tantamount to arguing in favor of the

[5] It can be easily proved by using the techniques of logical deduction that each of the interpretations from (i) to (iii) logically implies interpretation (iv).

fifth possibility that altruism is neither necessary nor sufficient for cooperation. But this very answer naturally leads us to an additional question: If *altruism* is neither necessary nor sufficient for cooperation, can it still be somehow connected to and act as a "contributory cause" of *cooperation*.

Now, this dissertation is mainly an attempt to explain, examine, and argue *against* or disprove the proposition, or rather hypothesis, that makes the weak claim that altruism is necessary or sufficient for cooperation to emerge. It is important to notice that the falsity of the fourth and weakest interpretation logically disproves each of the three preceding interpretations.[6] However, the contention that altruism is neither necessary nor sufficient for cooperation may be regarded as an answer to the central research question, "Is altruistic behavior by each member of a group of rational agents towards his or her opponent a necessary or a sufficient condition for mutual cooperation among them when their interests are mixed, i.e., partly common and partly conflicting?" By arguing against the weak hypothesis I will in effect try to defend the rather counterintuitive but factually *strong statement*[7] that altruistic behavior of everyone toward everyone else is neither a necessary nor a sufficient condition for mutual cooperation to emerge among rational individuals whose interests are neither entirely identical where mutual cooperation can easily happen nor diametrically opposed where it is absolutely impossible for cooperation to take place, but rather mixed which leave open the possibility of cooperation as well as defection.

[6] This is an intuitively evident implication that can be proved to be valid by using a modest knowledge of symbolic logic.

[7] The word "strong" is a relative term and given any two statements, one of them is factually stronger than the other when the former implies but is not implied by the latter. Thus, for example, the statement "Mary is a mother." is stronger than the statement "Mary is a woman." because the former obviously implies but is not implied by the latter. Moreover, denying a weak statement entails asserting a strong statement.

Thus my conception about the supposed causal connection between altruism and cooperation constitutes a fifth interpretation. For a reasonable defense of my view I will formulate two different arguments from counterexample on the basis of clarification of relevant concepts and utilization of empirical evidence available from secondary data. One argument from counterexample will be built to argue that cooperation may happen among egoists, while another argument from counterexample will be constructed to argue that cooperation may fail to occur even among altruists. Thus I will use the two different arguments to defend the view that cooperation may occur among egoists but may fail to occur among altruists. But this is equivalent to showing that altruism is neither a necessary nor a sufficient condition for cooperation to occur among a number of agents. This does not, however, imply that altruism and cooperation are not related in any other possible ways. Thus the thesis statement turns out that altruism is neither necessary nor sufficient for, but can still contribute to, cooperation just in virtue of being a *relevant condition*. An adequate defense of the thesis statement, however, would require not only putting forward arguments or counterarguments mentioned above but also defining the key as well as related terms, such as altruism, egoism, cooperation, mutualism, conflict, necessary condition, sufficient condition, relevant condition, rationality, social dilemmas, the prisoner's dilemma and the chicken as two different forms of social dilemma, and critically examining the related issues, e.g., the egoism-altruism debate.

1.4 Structure of the Dissertation

Having introduced the main research problem and the rationale for the research, I present here a brief outline of how the rest of the dissertation is organized.

Chapter 2 provides a comprehensive review of the relevant literature and conceptual framework of the study. Section 2.2 briefly explains and evaluates the rational choice theory that tries to base macro-behavior of humans on a micro-foundation. Section 2.3 discusses the problem of rationality which is related to the rational choice theory and deals with whether humans are rational or not. Subsections 2.3.1, 2.3.2, and 2.3.3 examine the various types of rationality, the nature and conditions of instrumental rationality, and the limits of instrumental rationality, respectively. Section 2.4 is divided into two subsections of which subsection 2.4.1 discusses the elements of game theory and subsection 2.4.2 discusses the nature and types of social dilemma games particularly the common good dilemma and the public good dilemma. Section 2.5 is divided into three subsections and deals with Darwin's theory of natural selection, its limits, and the Darwinian puzzle as a background to the theories of cooperation. Section 2.6 is divided into seven subsections and discusses seven different mechanisms of cooperation, viz., Kin Selection, Group (or Multilevel) Selection, Spatial Selection, Direct Reciprocity, Indirect Reciprocity, Strong Reciprocity, and Costly Signaling. Section 2.7 briefly examines the Egoism-Altruism Debate. Section 2.8 is divided into two subsections of which subsection 2.8.1 discusses the nature and types of conflict and subsection 2.8.2 discusses the nature and types of cooperation.

Chapter 3 is devoted to the discussion of methodological issues. Subsection 3.2 distinguishes between proof and disproof and then between two types of proof direct proof and indirect proof. Subsection 3.3 distinguishes among the three related concepts of conditions, causation, and correlation in order to shed light on the concept of cause.

In Chapter 4 the subsection 4.2 examines the Prisoner's Dilemma as a matrix (or strategic) form game and several strategies such as, All-D, All-C, and TFT, for dealing with numerous real life situations of conflict and cooperation. Subsection 4.2 examines the Chicken (or Hawk-Dove) game and presents it with the usual example of egoistic players and then with a novel example of altruistic players. Subsections 4.4 and 4.5 discuss the similarities and the differences between the two games, respectively in the light of the concepts of dominant strategy equilibrium, Nash equilibrium, correlated equilibrium, and Pareto optimality.

Chapter 5 may be regarded as the culmination of this dissertation in the sense that the *indirect proofs of validity* for two arguments of which one refutes the view that altruism is a necessary condition for cooperation and the other refutes that altruism is a sufficient condition for cooperation. Then we argue that just as smoking is neither necessary nor sufficient but is still a contributing condition for lung cancer, so also altruism is neither necessary nor sufficient but is yet a contributing condition for cooperation.

The conclusion of the dissertation is Chapter 6 which has three functions, viz., presenting a brief summary of the study, making a list of the main conclusions of the study, and offering suggestions for further research.

CHAPTER TWO

LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK

2.1 Preview

This chapter provides a comprehensive literature review and conceptual framework relevant for addressing the main research question as to whether *altruism* is only a *necessary* condition, or else only a *sufficient* condition, or *both* a necessary and sufficient condition, or else *neither* a necessary nor a sufficient condition but is

nevertheless a *contributing* condition for mutual *cooperation* among *rational* agents interacting in circumstances where their interests are neither entirely identical nor completely contradictory, but rather *mixed* that leave open the possibility of cooperation as well as defection. Based on this question, the discussion on literature review and conceptual framework is arranged under seven different sections, i.e., from section 2.2 to section 2.8, of this chapter that presents the background information on several key concepts, theories, models, issues, and disputes that are pertinent to the dissertation. Since our present knowledge about what causes cooperation is the product of cumulative efforts of previous scholars, our purpose here is to critically evaluate the existing research and debates relevant to the topic of our study and to identify inconsistencies and inadequacies in those studies, thereby making a case for why further study is required. The main focus is on the theories of the evolution of cooperation and a number of related topics such as the nature of rational choice theory, the distinction between different types of rationality with a special highlight on the nature, conditions, and limits of instrumental rationality, a brief introduction to game theory including social dilemmas (SD) as public goods games and common goods games, the prisoner's dilemma and the chicken (or, hawk-dove game) as two different forms of social dilemma, the theory of natural selection and the Darwinian puzzle, several important mechanisms for the development of cooperation, the distinction between biological and psychological altruism, the egoism-altruism debate, the meaning, nature, and types of conflict, the meaning, nature, and types of cooperation, and the distinction between cooperation, altruism, and mutualism.

2.2 Rational Choice Theory

Rational Choice Theory (RCT) is based on the concept of "rationality" and reflects the western values of individual liberty and equality. It attempts to explain the micro-macro relationship in economic, political, and social behavior by basing macro-behavior on a micro-foundation.

RCT deals with *how* a social or *collective choice* can be made *through an aggregation* of the *individual choices*. It involves a number of key characteristics. First, RCT is methodologically *individualist* in so far as it essentially begins with and gives priority to the desires, decisions, and deeds of the individual as opposed to those of the group. Secondly, RCT assumes that each individual is *instrumentally rational* which means that an individual's choice of action is *optimal* given the personal preferences, the opportunities available, and the constraints under which the choice is made. Optimality requires him to be able to make a *complete* and *consistent* ranking of his preferences over all the available options that are mutually exclusive and collectively exhaustive. Thirdly, the actions of the individual show *self-regard* in so far as they are concerned wholly with the pursuit of his own welfare or self-interests. Fourthly, RCT requires finding a level of *aggregate social welfare* out of preference rankings of all the individuals who are presumed to be rational beings seeking to maximize their respective benefit or welfare. Fifthly, it requires finding an effective and reasonable *method of amalgamation* for deriving a group preference ranking and thereby a collective choice out of the preference rankings of all the individuals.

Thus rational choice theory is an attempt to explain such social phenomena as choosing a policy or a code of conduct for an organization as an outcome of the preferences of its members who are assumed to be acting rationally at the individual level (Coleman, 1990; Hechter and Kanazawa, 1997). Social choices are usually made by different methods in different societies, e.g., various types of *voting* in democratic societies, *market mechanism* in market economies, and social *customs* and religious codes in traditional societies (Arrow, 1963). Thus RCT may be understood in terms of the key concepts of *individualism*, *optimality* or *rationality*, *self-regard*, *aggregate social welfare*, and an acceptable *method of amalgamation* of individual choices into a social choice.

Rational choice theory (RCT), as mentioned before, is based on the primacy of individual liberty and choice and tries to base collective choice on individual choices in an attempt to provide a micro-foundation of social choice. Now, RCT involves *methodological individualism* as opposed to *methodological collectivism*. Elster (1985) who tries to make sense of Marx in terms of methodological individualism and rational choice theory defines methodological individualism as "… the doctrine that all social phenomena – their structure and their change – are in principle explicable in ways that only involve individuals – their properties, their goals, their beliefs and their actions" (p. 5).

Methodological individualism, introduced by Max Weber in his *Economy and Society* (1922), makes two claims. First, it denies the existence of social entities independent of individuals. Secondly, it claims that explanation of social phenomena only in terms of causal connections or working regularities among social entities is not

enough and must ultimately appeal to an account of the activities at the level of the individuals. In other words, macro-analysis of social phenomena must be complemented by a micro-level analysis of behavior. Economists usually subscribe to the doctrine of methodological individualism. Even Adam Smith, the father of economics, may be regarded as a methodological individualist, as he, concerning the individual, wrote:

> He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. … he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. … By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it (Adam Smith, *The Wealth of Nations*, 1776, Bk. 4, Ch. 2.).

Methodological collectivism, as opposed to methodological individualism, was developed by Comte and Durkheim and supported by Hegelians and Marxists consider 'wholes' like 'society', an 'economy', 'capitalism', an 'industry', 'class', or a 'country' as collective phenomena that are given and hold that they are prior to facts about individuals from the explanatory point of view. According to methodological collectivism, the group exists and the independent and isolated existence of the individual is not conceivable, because he acquires his language, reason, beliefs, and customs which are essential to life from the society of which he is a member. Thus a methodological collectivist, as (Mises 1996: 42) points out, would conclude:

> "As the whole is both logically and temporally prior to its parts or members, the study of the individual is posterior to the study of society. The only adequate method for the scientific treatment of human problems is the method of universalism or collectivism." (p. 42)

Now, it is undeniable that all observable actions are actually performed by individuals and not by society which is only an abstract entity. There is, however, a sense in which we meaningfully say that groups also perform acts. Thus, we can only see an executioner and not the state that carries out a death sentence. But if we carefully examine the meanings of the various actions with reference to the respective contexts we can still significantly assert that the state has executed a criminal. On the other hand, the use of language, reason, and cooperation by an individual has no meaning apart from a society of which he is an integral part. Thus both individualism and collectivism overemphasizes only one side of the methodological issue and neglects the other side. As L. von Mises (1996) rightly points out: "Now the controversy whether the whole or its parts are logically prior is vain. Logically the notions of a whole and its parts are correlative. As logical concepts they are both apart from time" (p. 42).

Thus it is reasonable to deem both the positions as extreme and only partially true, and so it would be realistic and more reasonable to take a neutral stand and adopt a moderate view on the issue. However, it is important to note here that the viability of methodological individualism as a social choice theory could be more critically evaluated with respect to the realism of the assumptions of the theory, viz., *individualism*, *rationality*, *self-regard*, the derivability of an *acceptable level of aggregate social welfare*, and finding an effective and reasonable *method of aggregation* for obtaining a collective choice out of the preference rankings of all the individuals. Our main concern, however, is an evaluation of RCT only with respect to the problem of *rationality* which is the cornerstone of this theory and is taken up in Section 2.3. Having pointed out our

limited interest in RCT, we now proceed on to consider the problem of rationality and a host of other associated concepts and issues.

2.3 Problem of Rationality

The main problem about rationality is an interdisciplinary debate over the fundamental question whether humans are rational or not. Rationality is a core cognitive capacity of agents that can perceive, think, desire, decide, prefer, and choose, and hence the rationality of agents is a fundamental premise in the optimization sciences

where agents' actions are aimed at optimizing some well-defined objective function. Economics, for example, postulates that consumers act to maximize ordinal utility or expected utility, while firms act to maximize profit or expected profit (Hammond, 1997). However, our choices, beliefs, and the other cognitive capabilities seldom happen to be rational and frequently turn out to be irrational. Hence, the assumption of rationality has been under attack from many different quarters and has emerged as a hotly debated issue that has implications across a variety of disciplines.

The idea of rationality has quite a long history and its use can be traced at least as far back as Aristotle (*Nicomachean Ethics*, Bk. 1, Ch. 13) who considered rationality as an essential and unique attribute of man which differentiates humans from beasts. But Russell (2009) raised an objection against the conception of man as a rational being by way of a satire:

> "Man is a rational animal – so at least I have been told. Throughout a long life, I have looked diligently for evidence in favour of this statement, but so far I have not had the good fortune to come across it, though I have searched in many countries spread over three continents." (p. 45).

But the argument behind this humorous comment does not succeed as it is based on an *equivocation* between defining *mankind* as rational and describing an *individual man* as rational. It would, therefore, not be wise to assume that Aristotle was unaware of the widespread phenomenon of irrational behavior of individual human beings. If so, then we must look for a reason behind Aristotle's attempt at conceiving man as a rational being. Using the distinction between *competence* and *performance* which was first introduced by Chomsky (1965, 1980) in linguistics, Stich (1999) pointed out that Aristotle's view may be interpreted as the claim that all normal human beings have a *competence* or potential to develop into a rational man whether or not that potential is materialized in actual performance or behavior of individuals. It, therefore, remains a possibility to spell out the characteristics of a potentially rational man and to consider the various circumstances under which an individual's actual behavior may fall short of the standards of such a man.

2.3.1 Meanings of Rationality

The term "rationality" originates from the Latin word "***ratio***" meaning "***reason***". Even though this etymological meaning does not amount to a complete definition of rationality, yet it provides a clue to one by suggesting that *rationality* is somehow related to *reason*. Thus Kahneman (2011, 411), the Nobel laureate psychologist qua economist,

points out that in everyday speech the term '*rational person*' certainly refers to one who is '*reasonable*', and a reasonable person is one with whom it is possible to reason, whose 'beliefs' are in accordance with 'reality' and whose 'preferences' are consistent with his "interests' and 'values'. Philosophers, however, often draw distinctions between different types of rationality. For example, a major theme in Max Weber's philosophy is to identify four types of rationality, viz., practical, theoretical, substantive, and formal, and to draw comparisons between them (Kalberg, 1980). Even contemporary scholars also use the term "rationality" in different senses in different contexts. We, therefore, present below a brief survey of the term's usage which may be helpful to clarify the various meanings.

Rationality as Reasoning

A general definition of rationality which is quite broad and reflects the etymology of the word is formulated by the *OED* (2001) as "The quality of *possessing* reason; the power of being able to *exercise* one's reason" (p. 220). Obviously, this definition has two different but related components. First, it points out that rationality refers to the faculty of "reason" which, unlike "instinct" or "emotion", is an acquired or *learned capacity* and can guide one who possesses it to think or decide correctly. Secondly, it points out that an agent in possession of reason can be expected to be able to *apply* it when needed in specific circumstances.

Rationality as Sense of Proportion

Like the above *OED* definition based on the etymological meaning, there is also a commonsense approach that defines rationality directly in terms of the words "ratio" and "proportion". Let us take a look at the meanings of these two terms in order to fully bring out the meaning of the commonsense approach. By a *ratio* is meant a relationship between two different numbers showing the number of times one number contains or is contained within the other (*The Oxford Paperback Dictionary, Thesaurus, and Wordpower Guide*, New Delhi: Oxford Univ. Press, 2001, p. 738). And by a *proportion* is meant either the equality of two ratios or a *comparative part* or share or percentage *of a larger amount* (*Collins Concise Dictionary Plus*, ed., Hanks, Patrick, London: William Collins Sons & Co. Ltd., 1989, p. 1033). Now suppose that a given recipe of lemonade requires us to mix lemon juice with cold water in the ratio of 1:3. If someone who understands this recipe wants to follow it to prepare lemonade using 4 cups of lemon juice, then he or she must mix it with exactly 12 cups, and not any other quantity, of cold water. This implies that if he or she is to be considered a rational person while all other things remain the same, he or she must understand that the two ratios 1:3 and 4:12 are *equal*.

Thus we could say that the proportion of lemon juice to lemonade is ¼ and the proportion of cold water to lemonade is ¾. In other words, the ratio of the two proportions, ¼ : ¾, is equal to the ratio 1 : 3. Having explained the meanings of the words "ratio" and "proportion", we can now easily see that *rationality* consists in an agent's ability to *maintain* a "sense of proportion" and *irrationality* in going "out of proportion" in beliefs and behavior. Thus, rationality could be taken as the characteristic of behavior

based on reason and calculation and aimed at avoiding absurdity or illogicality. For instance, if a person regards his highest priority as passing an examination but allocates relatively more time to listening to music and participation in political activities than to preparations for the examination, his behavior could obviously be called irrational.

Theoretical vs. Practical Rationality

Traditional philosophical discourses draw a distinction between theoretical and practical rationality which has been succinctly stated by Svavarsdottir (2008) as follows: "Theoretical rationality is displayed in regulating one's beliefs and, perhaps, other cognitive states, while practical rationality is displayed in regulating one's actions, plans, intentions and, perhaps, other conative and affective states" (p. 1). Mele and Rawling (2004) also make the distinction in a similar way: "Whereas theoretical … rationality is concerned with what it is rational to believe, and sometimes with rational degrees of belief, practical rationality is concerned with what it is rational to do, or intend or desire to do" (p. 3). This distinction is also embedded in the general definition of rationality offered by *OED*.

Implicit in *OED*'s definition of rationality is the concept of theoretical rationality which consists in the possession of reason or general rules that make it theoretically possible for an agent to think, decide, or solve problems correctly. Thus, for example, if someone *understands* the *formula* "$a^n = a \times a \times a \times \ldots \times a$ (n times)", then it is *theoretically* possible that that person should be able to compute the value of 2^5 as well as that of 5^{660}, and hence may be said to be in possession of theoretical rationality. The reason for this is that theoretical rationality is based on the concept of theoretical

possibility which obviously does not admit of any degree. Hence, something must be either theoretically possible or impossible. The concept of theoretical rationality is particularly important in logic, mathematics, epistemology, and philosophy of science.

Also implicit in *OED*'s definition of rationality is the concept of practical rationality which consists in the capacity to *exercise* an agent's reason or general rule to think, decide, or solve problems correctly. Thus, for example, if someone *understands* the *formula* "$a^n = a \times a \times a \times \ldots \times a$ (n times)", then it may be *practically possible* for that person to compute the value of 2^5 but *practically impossible* to find the value of 5^{660}, and hence that person may be said to be in possession of limited practical rationality. The reason for this mismatch is that practical rationality is based on the concept of practical possibility which obviously admits of degree. Hence, whether something will be practically possible or impossible may be contingent upon many different conditions. The concept of practical rationality is of special importance in ethics, economics, political science, and sociology.

The reason for the discrepancy between the *possession* of reason and the *exercise* of one's reason is that the former does not guarantee the latter. Obviously, practical possibility logically implies but is not necessarily implied by theoretical possibility. This means that the concept of practical possibility is stronger, i.e., contains more information, than the concept of theoretical possibility. Hence, an important task for research in rationality is to discover or identify those contingent conditions that lead to the incompatibility between theoretical and practical possibility.

Epistemic vs. Instrumental Rationality

For the sake of convenience in focusing our attention on the meanings of epistemic and instrumental rationality and the distinction between them, let us begin by defining the two concepts in general terms. Epistemic rationality may be broadly understood as an attempt to examine *why* you *believe* what you believe. Instrumental rationality, on the other hand, may be generally understood as an attempt to make sense of what is the best you can do to get what you want to get.

Kelly (2003) defines the two terms 'epistemic rationality' and 'instrumental rationality' as follows:

> "By epistemic rationality, I mean, roughly, the kind of rationality which one displays when one believes propositions that are strongly supported by one's evidence and refrains from believing propositions that are improbable given one's evidence. … By instrumental rationality, I mean the rationality which one displays in taking the means to one's ends" (p. 612).

To put it in a little bit more technical terms but yet concisely, *epistemic rationality* may be defined as an agent's attempt at achieving knowledge, i.e., justified true belief, and avoiding false and unjustified beliefs by using accurate evidence and logically correct reasoning, while *instrumental rationality* may be defined as an agent's choice of the optimal, i.e., best possible, means to achieve a given end (Baron, 2008). But having a chosen end requires that the chooser be able to have a complete and consistent ranking of all the possible mutually exclusive options.

It may be pointed out that there is a dispute over whether or not epistemic rationality is a special case of instrumental rationality. While Kelly (2003) argues against the view that epistemic rationality is a special case of instrumental rationality, Leite (2007) takes the opposite stand on the issue. As the resolution of the debate does not affect the line of our argument, we will not examine the relevant arguments but would simply mention that we subscribe to the view that epistemic rationality is a kind of instrumental rationality. Epistemic rationality can be seen as a kind of instrumental rationality in which knowledge and truth are ends in themselves of which the optimal means are the best possible use of empirical evidence and logically valid arguments, while in a non-epistemic sort of instrumental rationality knowledge and truth may be instrumentally good for achieving the given goals. We will return to the topic of instrumental rationality in subsections **2.3.2** and **2.3.3** for discussion on its nature, conditions, and limits.

Individual vs. Collective Rationality

A useful theory of human action must be based on the assumption of stable and rational behavior that consists in an agent's having some goal or end and prudently choosing the means directed toward the achievement of the given end. On the basis of the nature of the social entity making the decision, Luce and Raiffa (1957) have divided choices into two kinds, viz., *individual choice*, and *group choice*. Now we normally observe the *individual* as a real decision-making unit to decide and act and evaluate those decisions and acts as displaying *individual rationality* or irrationality. But we also talk of

a group such as a club, firm or other organization consisting of different members as making decisions and appraise its acts as rational or irrational.

An individual and a group which is obviously more than a mere a collection of some individuals are not the same type of entity so that applying the attribute of rationality in the same sense to these two different kinds of entity would be, to use a well-known term from Ryle (1949), a 'category mistake'. Even if we deem the group as a functional as opposed to a biological entity, a difficulty arises when we thoroughly examine the question whether a group can be truly treated as a decision-making unit so that its so-called acts may be properly regarded as rational or irrational by the stipulated standards of rationality that should apply equally to an individual as well as a group. Unless we assume an organic conception of the social group, it is difficult to see why the group consisting of members of divergent interests, information, and expectations should decide upon a specific goal and why they should aim at achieving that goal.

However, if we postulate a set of goals commonly shared by all the members of the group, we may define collective rationality as the choice of a collective action which is consistent with the attainment of these goals (Buchanan & Tullock, 1962). It must be noted here that Buchanan and Tullock (1962) make an important distinction between individual and collective action, and thereby make the corresponding distinction between *individual* and *collective rationality*. An *individual action*, such as casting one's vote for any one of several candidates, is performed by a single or distinct individual who can be observed and recognized as a real decision-making unit and whose choice of action is based on a personal preference ordering over all the possible options. On the contrary, a *collective action*, such as the choice of a policy by the members of an organization, is not

any kind of action taken by a group as a whole but rather individual actions somehow aggregated to produce an outcome that is assumed to be commonly shared by all the members of the group. Thus though a collective action is only an aggregate of individual actions, yet the former unlike the latter does not have ordering of preferences. For example, riots, revolts, and other mob activities are mere aggregates of activities performed by rational, individual actors each of whom has a ranking of preferences.

Having discussed the distinction between individual and collective rationality, it is crucial but disappointing to note that these two types of rationality are conflicting rather than coherent. Obviously, the reason for conflict between individual and collective rationality is that what is good for society is not necessarily the same as what is good for each individual. The conflict between these two kinds of rationality which arises in the two well-known social dilemma games – the *Prisoner's Dilemma* and the *Chicken* – is discussed in chapter 4. Here we briefly illustrate the conflict with what is known as the *paradox of voting*.

The paradox of voting arises in connection with arriving at a social choice out of individual choices. There are several methods of making social choice, such as convention, dictatorship, voting, and market mechanism (Arrow, 1963). Social choice by ideal convention or ideal dictatorship involves no risk of clash of individual wills, and so can be definitely rational through satisfaction of the consistency or transitivity requirement of rationality. But then a question can be raised as to whether such rationality can be attributed to the method of voting which is a democratic system of making a collective political or social choice by aggregating many individual choices.

The the *paradox of voting* independently discovered by Condorcet (1785), Lewis Carroll and Edward J. Nanson, popularized by Duncan Black (See Riker, 1982, p.2), and generalized by Arrow (1963) as the famous Arrow's Impossibility Theorem is an attempt to answer the above question and shows that there can be no general procedure for amalgamating multiple individual rational choices into a collective social choice that would simultaneously satisfy the democratic norm of *majority principle*, on the one hand, and the *transitivity* condition of consistent or rational choice, on the other.

By the *majority principle* is meant a social decision rule such that given any two alternative objects of choice, x and y, x is socially preferred to y, i.e., x is chosen by society if and only if a majority of the voters prefer x to y. The *transitivity* principle which will be further explained and examined in the next two subsections and is required to ensure acyclic and well-ordered preferences, on the other hand, refers to the property of a binary relation such as preference, whereby an agent who confronts a choice among any three alternatives *x*, *y*, and *z*, prefers *x* to *y*, and prefers *y* to *z*, must also prefer *x* to *z* and not *z* to *x*. Now, to see that the paradox of voting shows that there is no method of amalgamating individual preferences into a collective choice which would simultaneously satisfy the *majority principle* and the *transitivity* principle, suppose that there are three candidates, A, B, and C, and three voters with their preferences over candidates arranged from left to right in order of decreasing intensity as shown by the following table:

TABLE 2.1: PARADOX OF VOTING			
	C A N D I D A T E S (From Left to Right in order of Decreasing Preference)		
VOTERS	1st Preference	2nd Preference	3rd Preference
Voter 1	A	B	C
Voter 2	B	C	A
Voter 3	C	A	B

As the table shows, voter 1 prefers A to B and B to C, and therefore by *transitivity* A to C. Voter 2 prefers B to C and C to A, and therefore by *transitivity* B to A. And voter 3 prefers C to A and A to B, and therefore by *transitivity* C to B. The table also shows that a majority including voter 1 and voter 3 prefer A to B, and a majority including voter 1 and voter 2 prefer B to C. It, therefore, logically follows by the *majority principle* that the *society* prefers A to B and B to C. But if the society is to be regarded as behaving rationally, i.e., according to the *transitivity* principle, it must prefer A to C. But in fact a majority of the society including voter 2 and voter 3, as shown by the table, prefer C to A. Therefore, by the *majority principle* society must prefer C to A. Thus by applying the transitivity principle and the majority principle to the given individual preferences in the table we arrive at a cyclic pattern of collective choice – A is preferred to B, B is preferred to C, and C is preferred to A – involving self-contradictions such as A is preferred to A, B is preferred to B, and C is preferred to C. The paradox of voting, therefore, shows the impossibility of any democratic method of amalgamating the rational preferences of three

or more individuals into a rational collective choice. And by utilizing symbolic logic, Alfred Tarski's (1901 – 1983) student Arrow (1963) took the result of this paradox to a higher level of generalization and sophistication through his "impossibility theorem" that shows in essence that there is no general mechanism of aggregating rational individual preferences over three or more alternatives that can satisfy several reasonable conditions of fairness and consistency such as unrestricted domain, Pareto optimality, independence of irrelevant alternatives, and non-dictatorship.

Descriptive, Normative, and Prescriptive Rationality

Corresponding to a traditional distinction in analytic philosophy between a *descriptive* (or, positive) *statement* and a *normative* (or, evaluative) *statement*, there are two different approaches – descriptive and normative – to understanding the nature of statements or judgments in various disciplines, and thereby dividing them between *descriptive science*, such as physics, chemistry, and biology, and *normative science*, such as ethics, aesthetics, and logic. Similarly, the term "rationality" which is used as an adjective to qualify many different cognitive capacities such as an agent's choices, preferences, decisions, expectations, behaviors, beliefs, and knowledge may be interpreted either in a *descriptive* sense to point out whether or to what extent those capacities are as a matter of fact guided by reason or in a *normative* sense to indicate that they ought on principle to be regulated by the use of reason (Hammond, 1997). Thus the descriptive approach to rationality belongs to psychology and refers to an *actual* state of affairs, i.e., the way agents decides and acts in real life situations, while the normative

approach to rationality belongs to mathematics and logic and refers to an *ideal* or desired state of affairs, i.e., the way agents should decide and act in ideal conditions of life.

The positive- normative distinction which is based on the Hume's (1739) question as to whether "ought "follows from "is" can be easily explained by examples (Lipsey, 1968; Beggs, 2020). Thus the statement

(1) The unemployment rate at present is 10 percent.

is a descriptive statement, as it communicates a factual information about the current state of the economy that may be tested through evidence. But the statements such as

(2) The unemployment rate at present is very high.

(3) The government ought to take action to decrease the unemployment rate.

are normative statements, as they are based on value judgments.

An essential feature of the descriptive-normative distinction is that it is logically impossible to deduce normative statements from descriptive statements and *vice versa*. It is important to note that, although the two normative statements (2) and (3) above are intuitively associated with the descriptive statement (1), they do not logically follow from the factual information provided in (1) unless they are combined with appropriate normative principles. Thus statements (2) and (3) need not necessarily be accepted given that the unemployment rate at present is 10 percent.

There is an important difference between positive statement and normative or value judgment about how a genuine, as opposed to a verbal or a logical, disagreement

may be resolved.[8] When there is a genuine dispute among reasonable persons about the truth or falsity of a positive statement, it may be resolved by an appeal to fact. But a real dispute among reasonable persons concerning a normative statement cannot be settled by an appeal to fact, because it involves a judgment about norms or values. A dispute about a value judgment is a not a *factual* disagreement but rather a disagreement in *attitude* which is extremely difficult, though not impossible, to resolve.

The fact that normative statements can be shown to contain some factual or descriptive elements and the positive-normative distinction often becomes blurred, has raised doubts about the possibility of, for example, a "value-free" positive science of economics (Sen, 1987; Putnam, 2002). Though descriptive statements are deemed as being permeated with normative values, it is worthwhile to retain the positive-normative distinction that contributes to clarity rather than confusion in decision and policy analysis (Lipsey, 1968; Weston, 1994).

There is, however, a gap between the *descriptive* and the *normative* approach to rationality, because actual human behavior, as will be discussed later on, often falls short of the requirements of ideal rational behavior. But this gap has been filled up by an important contribution of Bell, Raiffa, and Tversky (1988) who made a distinction between the *normative* and the *prescriptive* approaches which philosophers failed to see clearly and used the two terms synonymously (French, 1986). Thus following Bell, Raiffa, and Tversky (1988), we can distinguish among three different approaches –

[8] See Copi and Cohen (2004) and Machlup (1965) for extended and illuminating discussion on the nature and causes of different types of disagreement.

normative, *descriptive*, and *prescriptive* – to rational behavior. Let us take a brief look at the distinctions among the three perspectives.

Normative rationality is concerned with how ideal people would make decisions under stipulated norms or standards of ideal rationality. There are several important models of normative rationality, such as the expected value (EV) model, the expected utility (EU) model of von Neumann and Morgenstern (1947), and the subjective expected utility (SEU) model of Savage (1954) in decision making under risk or uncertainty, and probability theory and Bayesian statistics in the field of beliefs.

Descriptive rationality is concerned with how people actually make decisions in real life situations. The reliability of a descriptive model is determined by how far its predictions correspond to actual choices of people in real life. A highly celebrated descriptive model of decision making under risk and uncertainty is the *Prospect Theory* which was developed by Kahneman and Tversky (1979) and later on a polished and improved version of the theory was developed by Tversky and Kahneman (1992). This model was presented as a critique of EU theory as a descriptive model of decision making under risk and reveals many of the ways in which real life behavior of people depart from the normative standards of the EU model.

Prescriptive rationality is concerned with what people should and can do. It may be said to occupy the middle ground between the *descriptive* and the *normative* approach in so far as it measures the deviation of actual from ideal behavior and prescribes actions or policies to help people make practical improvements in moving closer to the normative ideal. The prescriptive perspective to rationality is characterized by a number of things,

such as: (1) helping people make better decisions by utilizing normative models, (2) recognition of the limited human capacity to make correct choice and judgment, (3) awareness of the practical problems of dealing with complex environments in which decisions are made, (4) the need for the suitable simplification of a complex condition of decision making, and (5) the requirement for training, tools, and debiasing techniques for utilizing a prescriptive decision model (Edwards, Miles, & von Winterfeldt, 2007).

Substantive vs. Procedural Rationality

Although both economics and psychology deal with human rationality, they had been doing it until recently in relative isolation and with emphasis on two different aspects of goal-directed behavior. While economics focuses on the *outcome* of a choice which is known as *substantive rationality*, psychology focuses on the *procedure* of a choice which is called *procedural rationality*. The distinction between substantive and procedural rationality has been extremely elegantly put by Simon (1976) as follows:

> The process of rational calculation is only interesting when it is non-trivial – that is, when the substantively rational response to a situation is not instantly obvious. If you put a quarter and a dime before a subject and tell him that he may have either one, but not both, it is easy to predict which he will choose, but not easy to learn anything about his cognitive processes. Hence, procedural rationality is usually studied in problem situations - situations in which the subject must gather information of various kinds and process it in different ways in order to arrive at a reasonable course of action, a solution to the problem (p.132).

Now, in the above example the set of mutually exclusive alternatives for choice is obviously fixed and given and the process of choice of an alternative on the basis of preference is simple. So, simply on the basis of the assumption of non-satiety, i.e., more is better than less, the agent would easily choose the quarter and reject the dime. Obviously, not all situations of choice are as simple as this one. Even in this simple case, studying the actual procedure of making the choice is much more difficult than merely understanding the logic of the choice. In more complicated cases, such as decision-making in business, the decision maker may be completely at dark about and may have to do a lot of hard work even to find out the alternative goals that may be available, to collect information about the alternatives in order to make up his or her mind about the relative desirability of them, to find out the realistically best means for achieving the goals, and so on. While a psychologist studying procedural rationality has to take into consideration all those difficult tasks about the process of decision making, an economist doing a 'logico-mathematical' study of substantive rationality ignores the complex problems about the decision making procedure and only concentrates on the whether the best or most preferred available alternative is chosen by making several ideal assumptions that all the possible alternatives are given, the decision maker is perfectly knowledgeable about the relative desirability of those options, and the goal of the agent is to pick the alternative that would provide him or her with the maximum benefit, such as maximum utility in case of a consumer and maximum profit in case of a firm.

2.3.2 Nature and Conditions of Instrumental Rationality

As defined earlier, instrumental rationality refers to a person's capacity to choose actions which can best satisfy his or her objectives which may, however, be either selfish or altruistic. Though instrumental rationality is a central concept in economics and the behavioral sciences, it has its root and a powerful defense in the philosophy of Hume (1740/1888). That rationality as goal-directed behavior is a common concept even in our ordinary life is evident from the following dialogue between Alice and the cat in *Alice's Adventures in Wonderland* which is one of the most popular works in English literature and written by Carroll (1865/2018):

> 'Would you tell me, please, which way I ought to go from here?'
>
> 'That depends a good deal on where you want to get to,' said the Cat.
>
> 'I don't much care where—' said Alice.
>
> 'Then it doesn't matter which way you go,' said the Cat (p. 52).

Hume's (1740/1888) claim that nothing but "passions" motivate a person to act and "reason" is only their servant is clearly expressed in his writing as follows:

> We speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be, the slave of the passions, and can never pretend to any other office than to serve and obey them (Treatise, Bk. II, Pt.III, Sec. III).

The standard normative ***model*** of instrumental rationality normally developed by economists consists of three broad elements (Mas-Collel, Whinston, & Green, 1995): First, it treats the decision maker's tastes as his or her basic characteristics and expresses them in the form of *preference relation*. Secondly, it imposes a set of *rationality axioms* on the decision maker's preferences. Thirdly, it analyzes the preference relation to deduce consequences that would determine the decision maker's *choice*.

We will now analyze the distinction first among the three related concepts – *choice*, *preference*, and *indifference*, then between the two concepts – *right choice* and *rational choice*, and then state the *conditions of rationality* that would guide an agent trying to make the *best choice*.

Life constantly compels us to make *choices*. And the quality of our life obviously depends on the value of the choices we make. According to the *Longman Dictionary of Contemporary English* (2000) to choose is "to decide which one of a number of things, possibilities, people etc that you want because it is the best or most suitable (p. 224)." A thorough analysis of this relatively short definition of the word "choose" would help us identify the following seven components of its meaning: (1) There must be an *agent* confronting circumstances where a choice must be made. (2) A number of *options* or alternatives are available which on further analysis can be seen to be mutually exclusive and collectively exhaustive and from which one has to be chosen. (3) A mental act of *preference* that involves *making up the mind* about the options on the basis of the *tastes* of the agent has to be performed. (4) An external act of *choice* that consists in *picking out an alternative* is naturally done after *making up the mind*. (5) The act of selecting one from among the available alternatives requires *rejecting the remaining alternatives* which

implies that the alternatives must be *mutually exclusive* and collectively exhaustive. (6) The selection is done for a *reason* which consists in the fact that *the one selected is the best* in the sense that it *optimizes* some value, such as utility, satisfaction, profit, or cost. It has to be noted that the option chosen must be on the top position in the preference hierarchy of the available options because of the assumption of *nonsatiety*, i.e., more is better than less, and hence the choice made must be the *best choice*. (7) If we stretch the definition of choice a little bit further, then it implicitly implies a time dimension because any choice is made at a particular point of *time*.

Having explained the different elements of the concept of choice, it would be the proper place to explain here, following Resnik's (1987, 12) distinction between a rational decision and a right decision, how a rational choice differs from a right choice. A rational choice is not the same as a right choice, though in most cases a rational choice tends to be the right choice. A *rational choice* is one made in accordance with the *appropriate rules* of decision-making and on the basis of *all relevant information* so that the choice is reasonably *expected* to lead to the *best possible outcome* in the given situation. A *right choice*, on the other hand, is one that luckily happens to be followed by the *best possible outcome* whether or not the decision was based on the appropriate rules and relevant information.

Suppose two persons – Tom and Dick – are both planning for an investment. On the prudent advice of a hired consultant Tom invests a big sum of money in the garment industry but due to unforeseen circumstances suffers a heavy loss. On the other hand, Dick decides simply on the flip of a coin not to invest his money in the garment industry but to put it in a bank, and thereby earns a considerable amount of interest. Here Tom's

choice was rational though it turned out to be a wrong choice, whereas Dick's choice was obviously irrational even though it luckily transpired to be the right choice. Obviously, we would like, if possible, always to make the right choice without ever bothering for whether or not the choice is rational. But unfortunately there is no guarantee for a choice as such or even a rational choice always to end up as the right choice. A rational choice will happen to be the right choice only in a situation where the agent is capable of making perfect calculation and has complete knowledge about the environment in which the choice is made. In fact, most of our decisions are made in uncertain or risky situation where we guess what *might* or is *likely* to happen, and hence we are uncertain whether or not such decisions will lead to the most desirable outcomes. Thus to sum up, all rational choices are not right choices, and *vice versa*.

Preference is a mental act of grading one thing as better than another which represents someone's tastes and underlies his or her act of choice. We may formally define ***preference*** as a binary relation between any two alternative things *x* and *y* such that someone prefers *x* to *y* if and only if that person has a greater desire for *x* over *y* during any given period of time. It has been characterized by Hansson & Grüne-Yanoff (2018) as *subjective comparative evaluations* of the form "Agent *A* prefers *x* to *y*". But we consider that it would be more accurate to describe preferences as *time-dependent subjective comparative evaluations* of the form "Agent *A* prefers *x* to *y* at time *t*". Firstly, preferences are *evaluations* in so far as they are concerned with value judgment about what should be chosen. Secondly, preferences are *subjective* in the sense that they are normally ascribed to agents. Thus, for example, even if most people consider that "*x* is better than *y*" on an objective basis, we may imagine two other persons of whom one

'prefers *x* to *y*' while the other 'prefers *y* to *x*' on equally reasonable grounds. Thirdly, preferences, as opposed to monadic concepts like "good" or "is desired", are *comparative* as they express the assessment of one thing *x relative to* another thing *y*. Fourthly, preferences are *time-dependent* in so far as they always happen in time, continue to exist over a period of time however short or long, and may cease to exist or the order of preferences may even reverse after that period.

We may now put the difference between preference and choice in a nutshell. Choice is an observable act of choosing something, while preference is a hidden psychological inclination for something against something else. Thus one's overt act of choosing coffee, when both tea and coffee are available, obviously reveals one's covert preference of coffee over tea. In a word, *choice* reveals and presupposes *preference*, but not conversely.

Like preference, ***indifference*** is also an evaluative binary relation that can be defined in terms of the former. An agent is *indifferent* between any two alternatives, *x* and *y,* such that after making a well-informed comparison between them she *prefers neither one to the other*. It is important not to confuse indifference with *ignorance*. Whereas preference implies a bias between the two options, both indifference and ignorance indicate an absence of one's psychological bias between them. There is, however, an *epistemological* as well as a *normative* distinction between them. One is said to be indifferent between two things, say, tea and coffee, if that person neither prefers tea to coffee nor coffee to tea. In other words one is *indifferent* between two things when they are either equally attractive or equally unattractive to her. *Ignorance,* on the other hand, is a situation where a person presented with a pair of options does not at all know

whether they are equally or unequally desirable, because she is unable to decide which option yields how much satisfaction or dissatisfaction.

Thus, for example, if a person pondering over the comparative desirability of watching a sunset scene by a beach and attending a seminar cannot make the evaluation, not because the two options are equally attractive but because she is unable to find out how much satisfaction she would get from each alternative, then she neither prefers one to the other nor is indifferent between them. She is rather ignorant about their relative merit or demerit. The basic difference is that a person is *indifferent* between any two options when she, on the basis of available information, ranks them on the same level, but is *ignorant* about them when she cannot judge them as being either equally or unequally desirable or undesirable. Thus indifference involves a value judgment based on subjective as well as objective consideration, but ignorance is devoid of any judgment.

It is now important to point out that an agent may have to make decisions under different kinds of environment. On the basis of whether and how much knowledge and information the decision-maker has about the environment, decisions or choices may be classified under three different kinds – decisions under *uncertainty*, *risk*, and *certainty* (Luce & Raiffa, 1957). You make your decision under *certainty*, if you know for certain which choice of action by you will lead to which outcome. You make your decision under *uncertainty*, if you have no information or idea about the outcomes of the various possible alternative acts. But you make your decision under *risk*, when you are in a state of neither total certainty nor total uncertainty but partial information, i.e., you know the degrees of probability for the outcomes of each possible alternative act and can utilize this information in making the decision. Now, the conditions of rationality that we are

going to sort out are appropriate for situations of decision making under *certainty* involving ideal agents who are perfectly well-informed about the alternatives and capable of making the most sophisticated calculation.

Now, rational behavior requires making a rational choice which, as mentioned earlier, consists in making the optimal or best choice. But whether an optimal choice exists or not depends on two things, viz., the alternatives available for choice and the possibility of ranking the alternatives in order of the preference relation (Bergin, 2005). To ensure the existence of optimal choices a set of assumptions regarding the alternatives as well as the preference relation is indispensable. Different scholars (Kreps, 1990; Mas-Collel, Whinston, & Green, 1995; Green, 2002; Hargreaves-Heap & Varoufakis, 2004; Bergin, 2005) have proposed slightly different axioms depending on whether a strong or a weak preference relation is taken as primitive. Though taking the weak preference as primitive has the advantage of logical simplicity, I prefer to use both the notions of strong preference and indifference because this is more in accord with everyday use of the word 'preference'. Of course, it might also be possible as well as logically elegant to take, like Kreps (1990), only strong preference as primitive and then use it to define weak preference and indifference.

Now, as a preliminary to stating the conditions of rational decision making, let us take a look at the logic of preference and then fix the meaning of a number of symbols. Suppose that a decision maker is presented with a pair of alternatives, *x* and *y*, and is asked how she compares them, i.e., whether x or y is better than the other in her judgment. If the answer is that '*x* is better than *y*', then we can rephrase it as '*x* is *strictly preferred* to *y*' and symbolize it as $x \succ y$. For each pair *x* and *y*, four logically possible

responses are possible: (1) *x* is better than *y*, but not the reverse. (2) *y* is better than *x*, but not the reverse. (3) Neither *x* nor *y* is better than the other. (4) *x* is better than *y* and *y* is better than *x*. Though the fourth response is a logical possibility, we must preclude it as it is incompatible with the semantics of ordinary language. It is now obvious that the third response which states that neither *x* is strictly preferred to *y* nor *y* is strictly preferred to *x* is a case of *indifference*, as defined earlier, between x and y. The first three responses can, therefore, be symbolized respectively as follows: (1) $x \succ y$, (2) $y \succ x$, and (3) $x \sim y$, or, equivalently as $[\neg(x \succ y) \wedge \neg(y \succ x)]$ where '$\succ$' is read as "*is strictly preferred to*", '$\sim$' as "*is indifferent to*", '$\neg$' as "*not*", and '$\wedge$' as "*and*". Add to these a few more symbols such as '$\veebar$' read as "*or but not and*" indicating exclusive 'or', '$\rightarrow$' as "*implies*", and $(\forall x)$ as "*For all values of x*".

We now list a set of conditions of rational behavior that can ensure a preference-ordering of the options and making the best choice as follows:

(1) Agents: Rational behavior implies that there is an agent confronting circumstances where she is capable of making a choice.

(2) Options: Choice also implies that an agent making a choice must sort out a set of available alternatives or options that are mutually exclusive and often collectively exhaustive.

(3) Irreflexivity of Strict Preference: $(\forall x) \neg(x \succ x)$. This means that there is no *x* such that $(x \succ x)$. In other words, no alternative is more desired than itself. In fact, this is the essential characteristic of strict preference.

(4) Asymmetry of Strict Preference: $(\forall x)\,(\forall y)\,[(x \succ y) \rightarrow \neg(y \succ x)]$. This means that given any two options x and y, if any one of them, say x, is strictly preferred to the other one, y, then y is not strictly preferred to x. In other words, preference reversal is not allowed for a given period of time.

(5) Transitivity of Strict Preference: $(\forall x)\,(\forall y)\,(\forall z)\,[\{(x \succ y) \wedge (y \succ z)\} \rightarrow (x \succ z)]$. This means that given any three options, if a first one is strictly preferred to a second and the second is strictly preferred to a third, then the first is strictly preferred to the third This condition is also known as the consistency condition and is imposed to ensure the consistency of choices and to rule out the possibility of cycles.

(6) Completeness: $(\forall x)\,(\forall y)\,[(x \succ y) \veebar (y \succ x) \veebar (x \sim y)]$. This means that given any pair of alternatives x and y, x is strictly preferred to y, or else y is strictly preferred to x, or else x is indifferent to y. In other words, for any pair of alternatives the agent is fully capable of comparing them and making up her mind as to whether she prefers one particular alternative to the other or else is indifferent between. If preferences were not complete, i.e., if an individual could not compare and rank all the alternatives, then it would not be possible for her to select the best alternative. It should be noted here that since the formal statement of the definition of completeness condition is a strong or exclusive disjunction where one of the three components, $(x \sim y)$, expresses an indifference relation which, though related to, is different from the preference relation, we include below explicit statement of the basic properties of indifference as additional conditions.

(7) Reflexivity of Indifference: $(\forall x)\,(x \sim x)$. This says given any option x it is indifferent to itself. Obviously, this property follows from the irreflexivity of strict preference which says that no alternative is more desired than itself.

(8) Symmetry of Indifference: $(\forall x)\,(\forall y)\,[(x \sim y) \rightarrow (y \sim x)]$. This says that given any two options x and y, if any one of them, say x, is indifferent to the other one, y, then y is also indifferent to x. The reason for this is simple, because it follows from the definition of indifference. If your desire for x is as great as your desire for y, then your desire for y is also as great as your desire for x.

(9) Transitivity of Indifference: $(\forall x)\,(\forall y)\,(\forall z)\,[\{(x \sim y) \wedge (y \sim z)\} \rightarrow (x \sim z)]$. This means that given any three options, if a first one is indifferent to a second and the second is indifferent to a third, then the first is indifferent to the third. The justification for this condition is that if each of two options, x and z, is as much liked as a third option y, then x is as much liked as z.

(10) Optimization: A rational agent's act of choice is purposive in so far as it is directed toward achieving a goal, i.e., optimizing an objective function, such as maximizing utility, maximizing profit, and minimizing cost. The individual's goal is to choose the most preferred alternative, i.e., the alternative that maximizes her net positive benefit.

2.3.3 Limits of Instrumental Rationality

Though the normative model of instrumental rationality is an important tool for judging behavior as either rational or irrational from a perfectly knowledgeable

perspective that does not admit of any distinction between different degrees of rationality, an appropriate application of it calls for the need to be aware of its limits. We organise our evaluation of the model under three different headings – unrealistic conditions, underestimation of the role of reason, and unexplained shift of ends – as follows:

Unrealistic Conditions

Standard theories of rational behaviour are based on the assumption that agents have hierarchically ordered and stable preferences that completely conform to the ordering conditions stated above. There are, however, arguments both for and against these conditions.

A number of scholars have pointed out that the consistency conditions of rationality are unrealistic because they are incompatible with either the observation of choices made in everyday life (Gauthier, 1986) or the laboratory experiments conducted on individuals whose choices revealed the phenomena of *intransitivity* and *preference reversal* (Tversky, 1969; Lichtenstein & Slovic, 1971; Tversky & Thaler, 1990). As to the difficulty of satisfying the *comparability* condition, let us imagine a hungry foreigner who is trying to choose an appropriate food for him from the menu in a hotel but fails. He may fail because of his inability to compare the various alternative items even after consulting a waiter and a great deal of pondering over the menu. His inability to compare the available alternatives is obviously due to his *unfamiliarity* with or *lack of definite knowledge* about them.

But many respectable scholars (Davis, 1958; Raiffa, 1968; Grether & Plott, 1979; Li, 2006; Regenwetter, Dana, & Davis-Stober, 2011) have seriously questioned the

soundness of the arguments against the consistency of preferences and pointed out the need for reinterpretation of the results of experiments as well as consideration of the implications of rejecting the consistency conditions. Thus, for example, Raiffa (1968) developed an ingenious thought experiment known as the "money-pump" argument to show that rational behavior must satisfy the transitivity of preferences. This argument is used to show that a series of *indifferences* combined with *intransitive* preferences can end in a real *difference*. Suppose, for example, that you prefer Coke to Sprite, Sprite to lemonade, and lemonade to Coke. Since you prefer lemonade to Coke, you would presumably be willing to offer more than the Coke to obtain the lemonade. Assume that you are ready to offer 1c. Thus you are ready to offer Coke+1c for lemonade. Furthermore, suppose that you are ready to offer lemonade+1c for Sprite, and Sprite+1c for Coke. Thus after three trades in the order just shown you find yourself with the initial Coke but 3c poorer. This implies that a clever trader can exploit your intransitive preferences to make you penniless by an endless round of such trades.

There is no doubt about the limits of the axiom of rationality as a foundation of economic theory. Human behaviour in everyday life often falls short of the ideal conditions of completeness and consistency due to a variety of factors, such as emotion, prejudice, laziness to engage in "slower, more deliberative, and more logical" thinking, and lack of adequate information, experience, or computational ability (Simon, 1976; Kahneman, 2011). But do the limitations of the assumption constitute a sufficient ground for rejecting it? Sen (1990) argued against the rejection:

> It will not be an easy task to find replacements for the standard assumptions of rational behaviour ... that can be found in the traditional economic literature, both because the identified deficiencies have been seen as calling for rather divergent remedies, and also

> because there is little hope of finding an alternative assumption structure that will be as simple and usable as the traditional assumptions of self-interest maximization, or of consistency of choice (p. 206).

Thus Sen's ground for not rejecting the conditions is the absence of an alternative set of assumptions that is at least as 'simple' and 'usable' as the prevailing conditions. To this we may add that there must be a gap, however big or small, between a genuine ideal and the actual. As soon as an actual practice catches up the ideal, the gap ceases to exist and so a new ideal must be set. We, therefore, consider that a divergence between the two is an essential precondition for improvement of practice. Hence, it is more reasonable to retain than to reject the traditional consistency condition of rationality, however unrealistic or unattainable.

Although, according to Hume's hypothesis, the choice of ends is left entirely to the agent himself, it would be useful for us to know his ends thoroughly in order to understand his behaviour properly. Conflicts may arise when a person sets up or pursues goals which are partially or wholly divergent from those of others. Hence, an important tip for the choice of ends could be that compatible goals must be preferable to incompatible ones. Let us take some examples from Russell (1954/1963):

> If a man and a woman desire to marry each other, both can be satisfied; but if two men desire to marry the same woman, one at least must be disappointed. If two partners both desire the prosperity of their firm, both can achieve the result; but if two rivals each desire to be richer than the other, one of them must fail. … When a nation is at war, the desires of all its citizens for victory are mutually compossible, but they are incompatible with the opposite desires of the enemy. The desires of those who feel benevolently to

> each other are compossible, but those who feel reciprocal malevolence have desires that are incompatible.
>
> It is obvious that there can be a greater total satisfaction of desire where desires are compossible than where they are incompatible. … It follows that love is preferable to hate, co-operation to competition, peace to war, and so on. (Of course, there are exceptions; …) (p.59).

Russell thus claims to have arrived at a standard by which he distinguishes between desires as right or wrong, or even as good or bad. As he puts it, "Right desires will be those that are capable of being compossible with as many other desires as possible; wrong desires will be those that can only be satisfied by thwarting other desires."

One of the most fundamental limitations of Hume's *means-ends model* is that it restricts rational behaviour to merely a choice among alternative means to a given end and cannot accommodate a choice among alternative ends. Consequently, it fails to explain why a given rational person may have to shift from one end to another.

Obviously, the *means-ends model* is included in the *preferences-opportunities model*, because the former is a special case of the latter where the person has settled for a given objective for which he is trying only to find the best means.

2.4 Game Theory and Social Dilemmas

The overall well-being of any society is threatened by the existence of social problems, such as overfishing in the rivers and oceans, overgrazing, overpopulation, environmental pollution, and so on, that arise out of the conflict between individuals'

short-term self-interests and society's long-term collective interests. Such problems are generally known as social dilemmas which are interdisciplinary in nature and better explained in terms of the mathematical theory of games. Knowledge of game theory and social dilemmas can help us better understand why sometimes people cooperate and why sometimes they do not. We first discuss game theory and then social dilemma games as follows:

2.4.1 Elements of Game Theory

Game theory provides powerful analytical tools for various disciplines and is rapidly flourishing as a field of study. Although the origins of modern game theory can be traced to the works of several earlier mathematicians, it was von Neumann and Morgenstern (1944) who did much of the groundbreaking work on this theory. In fact, more than a dozen of reputed scholars particularly John Nash, John Harsanyi, Reinhard Selten, Thomas Schelling, and Robert Aumann have so far been awarded Nobel prize in economics for their substantial contributions in game theory.

Let us begin by pointing out a distinction that is often made between *game theory* and *decision theory*. *Decision theory* is the study of how a single individual can maximize her or his expected benefit through the choice of an action from a given number of alternative courses of action. For example, an individual's choice of an *act* as to whether she or he should *carry* or *not carry* an umbrella while going out of doors based on two different things, viz., *states of nature* (here, the weather condition, such as raining and not raining) and her or his *utility* or *preference function* for ranking all the

possible alternative outcomes (such as, take umbrella and it rains, take umbrella and it doesn't rain, don't take umbrella and it rains, don't take umbrella and it doesn't rain) determined by all the possible combinations of acts of the decision-maker and states of nature. Obviously, the best outcome for someone in these circumstances would be not taking the umbrella while it does not rain and the worst one would be not taking the umbrella while it does rain. Now, decision theory is concerned with just a single-person choice known as a *game against nature* where nature is considered as a disinterested player.

There are three *branches* of decision theory, viz., descriptive decision theory, normative decision theory, and prescriptive decision theory. *Descriptive decision theory* studies how actual human beings, rational or irrational, make decisions. *Normative decision theory* considers how perfectly rational beings would make optimal decisions. *Prescriptive decision theory*, however, tries to provide guidelines for agents to make the best possible decisions given the fact that men make mistakes in making up their minds in real life and are endowed only with *bounded rationality* due to the limits to their thinking capacity, available information, and time (Simon, 1987). Decision theory framework, as pointed out earlier in this chapter, generally deals with three *types of decision*, viz., decisions under *certainty*, decisions under *uncertainty*, and decisions under *risk*.

Let us now define game and then game theory. A *game* may be defined as any situation of interaction between two or more intelligent and goal-oriented players each facing a choice between two or more alternative acts but the outcome depends on the choices of all the players. Thus a game has the following components:

(1) There are at least two ***players*** or agents who make a decision.

(2) Each player is assumed to be a ***rational*** agent capable of consistently pursuing her goals.

(3) Each player has a number of possible strategies or contingent courses of action to choose from. A ***strategy*** is a complete "plan of action" which refers to a set of contingent instructions or conditional actions, such as 'Do A if X occurs, and do B otherwise'.

(4) For any interaction or game between two or more players there is a number of possible ***outcomes*** each jointly determined by a set of strategies chosen, one by each player.

(5) Each possible outcome of the game is composed of a collection of numerical ***payoffs***, one for each player, which show the monetary values or utilities of the outcome to the different players.

(6) Each player has access to the ***information*** about the entire situation at each decision time.

Now, *game theory* is the scientific study of games. Its main purpose is to find an appropriate *solution* of a game i.e., to find out an outcome where all the players involved would or could settle. Myerson (2013) has defined game theory as "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers (p. 1)." According to Colman (2005), "Game theory is a formal theory of interactive decision making, used to model any decision involving two or more decision makers, called players, each with two or more ways of acting, called strategies, and well-

defined preferences among the possible outcomes, represented by numerical payoffs" (p. 688).

There are two fundamental characteristics of game theory, viz., instrumental rationality and common knowledge of rationality (Hargreaves-Heap & Varoufakis, 2004). First, all the players are usually assumed to be ***instrumentally rational*** in the sense of having an *end* and choosing the best available *means* to that end. This means that every player has well-defined preferences, beliefs about the decision-making environment including the other players, and tries to optimize his or her personal payoffs. Second, players have ***common knowledge of rationality*** and they know the rules of the game. This means that each rational player is aware that other players are also rational and trying to optimize their payoffs.

Now, according to different criteria, games may be classified into various contrasting categories (Colman, 1982, 1999; Fudenberg & Tirole, 1991; Hargreaves-Heap & Varoufakis, 2004; Prisner, 2014), such as (1) Games of Skill, Games of Chance, and Games of Strategy (based on the factors that influence the outcomes of the games), (2) Two-person Game vs. Multi-person (or, N-person) Game, (3) Interactive vs. Non-interactive Games (or, Games against Nature), (4) Constant-sum vs. Non-constant-sum Games, (5) Zero-sum vs. Non-zero-sum Games, (6) Strictly Competitive Games, Pure Coordination Games, and Mixed-Motive Games, (7) Cooperative vs. Non-cooperative Games, (8) Symmetric vs. Asymmetric Games, (9) Simultaneous Move Games vs. Sequential Move Games, (10) Matrix (or Normal, or Strategic) Form Games vs. Extensive (or Dynamic, or Tree) Form Games, (11) One-shot Games vs. Repeated Games, (12) Complete Information vs. Incomplete Information Games, (13) Perfect

Information vs. Imperfect Information Games, (14) Deterministic vs. Stochastic Games, (15) Static vs. Dynamic (or, Evolutionary Games), and (16) Discrete vs. Continuous Games.

It must be noted that the above classification of games into different categories are based on different criteria, and hence one category of game based on some criterion may be overlapping with another category of game based on a different criterion. We will focus on the division of games into Strictly Competitive Games, Pure Coordination Games, and Mixed-Motive Games. This classification is founded on one important feature of games, viz., how the payoffs or interests of the players are related. The interests of the players making independent but interdependent decisions may be related in any one of three possible ways. The players' interests as reflected in the outcomes of a game may be *completely conflicting*, or *entirely identical*, or *mixed*, i.e., partly conflicting and partly common. We will now discuss three different games, viz., the ***Penny-Matching Game***, the ***Driving Game***, and the ***Disarmament Game***, each of which represents one of the three categories. But as an aid to explaining these games and related concepts and principles, we will also discuss several other games and matrices including, for example, the famous ***Stag-Hunt Game***.

Let us first take up the ***Penny-Matching Game*** which is an example of a basic game involving two players – Player I and Player II – who are rational decision-makers seeking to maximize their respective interests or payoffs but whose interests are diametrically opposed as depicted in Table 2.2. In this game the two players, I and II, *simultaneously* place a penny on the table where the payoff depends on whether the pennies match or not. If they match, that is, both pennies are heads or both pennies are

tails, then one of the players, Player I by convention, wins and gets the other player's penny; but if they do not match, then Player II wins and gets the other player's penny.

TABLE 2.2: PENNY-MATCHING GAME

(A Zero-sum Game)

	Player II		
Player I	Strategy	Heads	Tails
	Heads	(+1, -1)	(-1, +1)
	Tails	(-1, +1)	(+1, -1)

As mentioned before, the *pay-off* for any player is the amount won or lost by that player, and an ordered pair of payoff numbers is called an *outcome* determined by the choices of the players in any given situation. Now, in Table 2.2 each of the four cells with a pair of numbers contains an *outcome* denoted by a specific pair of ordered numbers where, by convention, the first number denotes the *payoff* to the row chooser, i.e., Player I, and the second number denotes the *payoff* to the column chooser, i.e., Player II. The number +1 indicates a gain of one penny and -1 indicates a loss of one penny.

The game depicted in Table 2.2 is called a *zero-sum game* which is a situation where one player gains exactly what the other loses so that the sum of the two players' payoffs in each cell is equal to zero and the net change in wealth or benefit is zero. Thus Matching the Pennies is a *two-person*, *two-strategy*, *simultaneous-move*, *zero-sum* game.

There is a wide variety of situations pertaining to economic, political, social, military, and interpersonal interactions that can be modeled as *strictly competitive games*. For *example*, indoor games such as, poker, gambling, hide-and-seek, and chess, outdoor games such as, tennis, and penalty kick game in football, two TV networks competing for viewers, two rival politicians campaigning for votes, and two armies fighting for land can be modeled as a two-person, zero-sum game.

It is important to note that Matching the Pennies being a zero-sum game is a ***strictly competitive game***, i.e., a win-lose or lose-win game. Since in such a game one player can gain only at the cost of the other, there are *no possibilities* of mutually beneficial *cooperation*. Matching the Pennies is a ***Non-cooperative Game***, as opposed to a *Cooperative Game*, in the sense that there is no possibility of a binding or *enforceable agreement* between the players about how the game should be played.

To explain how best to play the *Penny-Matching Game*, and for that matter, any other game, it is crucial to define a few more extremely important concepts of game theory such as, Dominant Strategy, Dominated Strategy, Dominance Principle, Dominant Strategy Equilibrium, Nash equilibrium, Mixed Strategy Nash equilibrium, and Pareto-optimality Principle.

As stated earlier, a *strategy* for a particular player is a complete plan of conditional action contingent upon what the other player or players are expected to do. Now, a given strategy compared to any other strategy of a player in a game must be related in any one of three different ways. Thus for any two strategies S_1 and S_2, either S_1 dominates S_2 so that S_2 is dominated by S_1, or S_2 dominates S_1 so that S_1 is dominated by

S_2, or neither S_1 nor S_2 dominates the other. It must be noted that there is a distinction between two types of dominant strategy, viz., strictly dominant strategy and weakly dominant strategy (See Hargreaves-Heap & Varoufakis, 2004). Given any two strategies S_1 and S_2 for any player, S_1 is ***strictly dominates*** S_2 if it guarantees him or her a *better* pay-off than any other alternative strategy would yield against all possible strategies of the opponent. If any strategy S_1 strictly dominates S_2, then S_1 is called a *strictly dominant strategy* and S_2 is called a *strictly dominated strategy*. But a strategy S_1 ***weakly dominates*** S_2, if S_1 guarantees a player *at least as good* a pay-off as S_2 would yield against all possible strategies of the opponent and a higher pay-off against at least one strategy of the rival player. If any strategy S_1 weakly dominates S_2, then S_1 is called a *weakly dominant strategy* and S_2 is called a *weakly dominated strategy*. From the concepts of dominant act and dominated acts follows what is called the ***dominance principle*** which holds that a rational player must eliminate all dominated strategies and choose the dominant strategy, if there remains one (Resnik, 1987; Straffin, 1993; Hargreaves-Heap & Varoufakis, 2004).

To illustrate the *dominance principle* and the related concepts, let us consider Table 2.3 that shows a pay-off matrix for a two-person non-zero-sum game where Player I has three strategies, R_1, R_2, and R_3, and Player II also has three strategies, C_1, C_2, and C_3. This table shows that for Player II, strategy C_3 is strictly dominated by strategy C_2 and strategy C_1 is weakly dominated by strategy C_2. Though in the original matrix there is no strategy for Player I that is dominated either strongly or weakly by any other strategy, in the reduced matrix when strategies C_1 and C_3 are eliminated in accordance with the dominance principle both the strategies R_1 and R_2 are strictly dominated by the

strategy R_3. Thus the successive elimination of all dominated strategies – C_1, C_3, R_1, and R_2 – by repeated use of the dominance principle leaves each player only with her dominant strategy, R_3 for Player I and C_2 for Player II which jointly determine the outcome (4, 2) which is what game theorists call the ***dominant strategy equilibrium*** that happens to be the optimal solution to this game[9].

TABLE 2.3: STRICT DOMINANCE vs. WEAK DOMINANCE

	Player II			
Player I	Strategies	C_1	C_2	C_3
	R_1	(3, 0)	(2, 1)	(0, 0)
	R_2	(1, 1)	(1, 1)	(5, 0)
	R_3	(0,1)	(4, 2)*	(0, 1)

But the dominant strategy equilibrium solution is not available for most games, because they do not have a dominant strategy for each player. And so we have to turn to a general and celebrated solution concept known as the Nash equilibrium. A game may have either a *Pure Strategy Nash Equilibrium* or *a Mixed Strategy Nash Equilibrium*. Nash (1950, 1951) proved that every finite non-cooperative game has at least one Nash equilibrium in mixed strategies. A *pure strategy* is one that is unconditionally used, i.e.,

[9] The use of the dominance principle for the elimination of strictly dominated strategies is useful for finding out any *dominant strategy equilibrium* that is necessarily a *Nash equilibrium*, but not *vice versa*. However, the application of this principle for the elimination of weakly dominated strategies can turn out to be problematic in so far as some other Nash equilibria may be lost in the process and the reduced game does not resemble the original one from a strategic point of view.

used with certainty or *100% probability*. A *mixed strategy*, on the other hand, is an assignment of a positive probability to each pure strategy such that the sum of the probabilities is 100% or 1. Thus a pure strategy is a special case of mixed strategy, because a player by definition chooses a pure strategy with certainty or 100% probability.

A ***Nash equilibrium*** is a solution to a non-cooperative game and may be defined as a *set of strategies*, one for each player, such that each player's choice of strategy is the *best reply* given the other player's choice of strategy so that no single player can obtain a higher pay-off by unilaterally switching to a different strategy. Now, a ***Pure Strategy Nash Equilibrium*** could defined as a as a *set of strategies*, one for each player, such that each player's chosen strategy is unconditionally played with a probability of 100%. A ***Mixed Strategy*** (i.e., ***Randomized***) ***Nash Equilibrium***, on the other hand, is one where each player mixes the different pure strategies with different degrees of probability so that the sum of the probabilities equals 100%.

We have seen that the Nash equilibrium is just an outcome where the rational players tend to settle at but says nothing about its desirability from the social perspective. We will now define and briefly discuss the *Pareto-optimality Principle* which originates from Vilfredo Pareto (1848-1923) and is used by economists to compare and evaluate any two alternative outcomes, allocations of resources, policies, systems, or states of affairs with respect to efficiency. According to the Pareto Principle, an outcome is ***Pareto-optimal*** (or, Pareto-efficient) if and only if there is no alternative outcome that would make at least one person better off without making anyone else worse off. This means that an allocation is ***Pareto-suboptimal*** (or, Pareto-inefficient) if and only if there

is at least one alternative outcome that would make at least one person better off without making anyone else worse off.

A distinction is usually drawn between two versions of the Pareto Principle, viz., the Strong Pareto Principle and the Weak Pareto Principle (See, for example, Johansson & Löfgren, 2003; Cato, 2013):

Strong Pareto Principle: If at least one individual prefers an outcome *x* to another outcome *y* and nobody prefers *y* to *x*, then society must prefer *x* to *y*.

Weak Pareto Principle: If every individual prefers an outcome *x* to another outcome *y*, then society must prefer *x* to *y*.

Having discussed the basics of the relevant theoretical and conceptual tools, let us now return to the problem raised earlier about how best to the play the game of *Matching the Pennies*. But let us first take a brief look at the game in Table 2.3. Here the outcome (4, 2) determined by the strategy profile (R_3, C_2) is not only a *dominant strategy equilibrium*, as shown earlier, but also a *Nash equilibrium*, because neither player can increase her payoff by choosing a different strategy. However, none of the eight other outcomes is a Nash equilibrium, because for each of these eight outcomes at least one player has an opportunity to increase her payoff by unilaterally switching to a different strategy. Moreover, outcome (4, 2) is also *Pareto-optimal*, because it is not possible to raise the payoff of one player without reducing the payoff of the other. It may be checked that no other outcome in Table 2.3 is Pareto-optimal simply because switching to any of the other outcomes yields a lower payoff for Player II.

To find the Nash Equilibria for the Penny-matching strategic game in Table 2.2, we now examine each strategy profile in turn.

(Head, Head):

Player II can raise her payoff from -1 to +1 by choosing the strategy *Tail* rather than the strategy *Head*. Thus the strategy profile (Head, Head) is not a Nash equilibrium.

(Head, Tail):

Player I can raise her payoff from -1 to +1 by choosing the strategy *Tail* rather than the strategy *Head*. Thus the strategy profile (Head, Tail) is not a Nash equilibrium.

(Tail, Head):

Player I can raise her payoff from -1 to +1 by choosing the strategy *Head* rather than the strategy *Tail*. Thus the strategy profile (Tail, Head) is not a Nash equilibrium.

(Tail, Tail):

Player II can raise her payoff from -1 to +1 by choosing the strategy *Head* rather than the strategy *Tail*. Thus the strategy profile (Tail, Tail) is not a Nash equilibrium.

It, therefore, follows that the game of Matching Pennies, as depicted by the pay-off matrix in Table 2.2, has *no Nash Equilibria* in pure strategies. Since both players in this game do not have a dominant strategy, there can be *no dominant strategy equilibrium* here. Had there been one, it would necessarily have to be a Nash equilibrium. Let us now consider if there is any Pareto-optimal outcome for this game. Just a quick glance at Table 2.2 shows that it is impossible to improve any one player's payoff without reducing the other player's pay-off by shifting from any outcome to any other outcome. Hence, there is *no Pareto-efficient* outcome of this game.

Although for Matching Pennies there are no Nash Equilibria in pure strategies, there must be, according to Nash (1950, 1951), one in mixed strategies. Let us now solve this game for the *Mixed Strategy Nash Equilibrium* (MSNE) and *expected payoffs* for the two players. A short computation for the solution values of this game has been shown in Table 2.4.

TABLE 2.4: COMPUTATION OF THE MIXED STRATEGY NASH EQUILIBRIUM & EXPECTED PAYOFFS FOR THE PENNY-MATCHING GAME

<table>
<tr><td></td><td colspan="3">Player II</td><td></td></tr>
<tr><td rowspan="3">Player I</td><td></td><td>Heads (p)</td><td>Tails (1–p)</td><td rowspan="3">Player I's Expected Payoff (EPI):
EPI(H)=+1p+(-1)(1-p)=2p-1
EPI(T)= -1p+1(1-p)=1-2p
Equating these EPs & Solving for p yields p=½.
EPI(H)= EPI(T)=1-2p=1-2(½)= 0</td></tr>
<tr><td>Heads (q)</td><td>(+1, -1)</td><td>(-1, +1)</td></tr>
<tr><td>Tails (1–q)</td><td>(-1, +1)</td><td>(+1, -1)</td></tr>
<tr><td colspan="4">Player II's Expected Payoff (EPII):
EPII(H)= -1q+1(1-q)=1-2q
EPII(T)=1q+(-1)(1-q)=2q-1
Equating these EPs & Solving for q yields q=½.
EPII(H)= EPII(T)= 2q-1=2(½)-1= 0</td><td>Solution:
MSNE Strategy Profiles
={(½H, ½T), (½H, ½T)}
MSNE Payoffs=(0, 0)</td></tr>
</table>

To get a clue to the solution, just assume that one of the players plays either Heads only or Tails only 100% of the times. Then the other player will soon come to know of this and exploit this information to defeat the opponent. If instead one player shows Heads only 60% of the times, the other player would also exploit this to outperform her rival. This suggests that to make the best response to the other each player must randomize between her two pure strategies by assigning equal probability for each strategy in order to make the rival indifferent between her own strategies.

Obviously, randomization can happen only when the expected payoffs for each player from her alternative strategies are equal. It is also important to note that each player's choice of strategy depends on the probabilities with which the other player plays her strategies. First, we assume that Player II plays Heads with probability *p* and hence Tails with probability (1-*p*), and Player I plays Heads with probability *q* and hence Tails with probability (1-*q*). Then we calculate each player's ***Expected Payoffs*** (EP) for Heads and Tails through multiplying her own relevant payoff numbers by the corresponding probabilities of the other player, form an equation of the EPs of each player, and solve for the values of *p* and *q* to find the probabilities with which each player should play her strategies as well as to find the EP values.

Having discussed the Penny-Matching Game, let us now consider the ***Driving Game*** depicted in Table 2.5 below as a situation where two persons are driving along a road from opposite directions. As the table shows, this is a two-person game where the Row Player which is Player I by convention is acting as Driver I and the Column Player which is Player II is acting as Driver II. Each player has two strategies, i.e., ways of acting – Driving on the Left or Driving on the Right. Player I's and Player II's choices, therefore, lead to four possible combinations of strategy choice – (Left, Left), (Left, Right), (Right, Left), and (Right, Right) – which respectively determine the outcomes (+1, +1), (-1, -1), (-1, -1), (+1, +1). Each outcome is, as usual, composed of an ordered pair of numerals, positive or negative, where the first numeral indicates Driver I's payoff and the second Driver II's payoff.

TABLE 2.5: DRIVING GAME (A Coordination Game)			
	Driver II		
Driver I	Strategy	Left	Right
	Left	+1, +1	-1, -1
	Right	-1, -1	+1, +1

Let us now check the four outcomes to find out which, if any, of the outcomes are *Pure Strategy Nash equilibria* (PSNE). We will also examine whether there is any Pareto-optimal outcome. Inspection shows that there are two PSNEs in the Driving game. These are (+1, +1) determined by the strategy profile (Left, Left) and again (+1, +1) by (Right, Right). The reason why they are PSNEs is that unilateral switch over from any of these two outcomes by any one player would reduce the payoff from +1 to -1. The outcomes (Left, Right) and (Right, Left) are not PSNEs. Each of the two *Nash Equilibrium* outcomes yielding the same payoffs to the two players, (+1, +1), is also *Pareto-optimal*, simply because no one can be made better off by a movement away from it.

Now, the Driving Game is a non-zero-sum game, because it has no outcome where the sum of the players' payoffs is equal to zero. In fact, a game to be non-zero-sum requires having at least one outcome where the sum of the players' payoffs is not equal to zero. The Driving Game is called a *pure coordination game* because it is a game in which all of the players' interests or preferences are identical, and hence there is no conflict of

interest between the players. In a coordination game, therefore, the players' only objective is to coordinate their strategies in a way that would help them achieve an outcome that they all prefer. A coordination game, however, is not trivial because experimental evidence has shown that coordination failure often happens in these games even when there is communication (Van Huyck, Battalio, & Beil, 1990; Cooper, DeJong, Forsythe, & Ross, 1990; Cooper, De Jong, Forsythe, & Ross, 1992; Pulford, Colman, & Lawrence, 2014; Dong, Montero, & Possajennikov, 2018).

That players may together fail to attain the outcome that is most preferred by all due to coordination failure is best understood from the two-person Stag-Hunt Game depicted by Table 2.6 below. The ***Stag Hunt Game*** originated from a short story that contains the essence of the social contract and is narrated by Rousseau in *A Discourse on Inequality*:

> "If it was a matter of hunting a deer, everyone well realized that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple..." (Cited in Skyrms, 2004, p. 1).

TABLE 2.6: STAG-HUNT (or, ASSURANCE) GAME

	Hunter II		
Hunter I	Strategy	Stag (C)	Hare (D)
	Stag (C)	3, 3	0, 2
	Hare (D)	2, 0	1, 1

The Stag-Hunt Game which is also known as "assurance game", "coordination game", and "trust dilemma" is a game which describes a conflict between *safety* (i.e., choosing Hare that ensures a minimum payoff of 1) and a much *better payoff*, 3, (i.e., choosing Stag that has the possibility of getting the worst payoff of 0) by social cooperation. As Table 2.6 shows two individuals go out on a hunt. Each hunter can choose to hunt a stag or hunt a hare. If one wants to succeed in hunting a stag, one must have the cooperation of one's partner. One can hunt a hare by oneself, but a hare is worth less than half a stag. It is better for both to jointly hunt a stag and share it than to individually hunt a hare. But unfortunately no one can trust the other because there is no assurance that the other will cooperate.

Checking Table 2.6 shows that this game has *no dominant strategy equilibrium*, but has *two PSNE*, (3, 3) and (1, 1), of which the former is *risk dominant* and the latter is *payoff dominant*. The strategy profile (Stag, Stag) is payoff dominant since payoffs are better for both players compared to the other strategy pair (Hare, Hare). On the other hand, the outcome (Hare, Hare) risk dominates the outcome (Stag, Stag), because there is uncertainty about the other player hunting a stag. Sufficiently high uncertainty about the other player's cooperation to hunt a stag will provide a hunter with a higher expected payoff from choosing 'Hare' than from choosing 'Stag'. The risky Stag-Stag solution is *Pareto-optimal*, while the safe Hare-Hare solution is *Pareto-suboptimal* Nash equilibrium.

We can find by appropriate calculation that the MSNE solution for the game requires the strategy profile to be {(½Stag, ½ Hare), (½Stag, ½Hare)} and the payoffs to be (1½, 1½). Thus the payoffs of the mixed strategy solution are better than those of the

low paying safe equilibrium but still much lower than those of the risky but Pareto-optimal Nash equilibrium outcome. Hence, the most important problem posed by the Stag-Hunt game is to find out how rational players can get from the safe but inefficient Hare-Hunting equilibrium to the risky but mutually advantageous Stag-Hunting equilibrium. An important lesson from the game, however, is that the mere existence of a mutually advantageous and efficient equilibrium outcome may not necessarily make it possible for the players to achieve it.

Having discussed the *Penny-Matching Game* and the *Driving Game* along with several other games as an aid to explanation, we will now take up the *Disarmament Game* as an example of the third general category of game known as the *Mixed-motive Games*. In a Mixed-motive Game the interests of the players are neither *wholly conflicting* nor *absolutely identical*, but rather *mixed*, i.e., partly conflicting and partly consistent. As Colman (1982) puts the distinction:

> Games in which the players' preferences among the outcomes are neither identical (as in pure coordination games) nor diametrically opposed (zero-sum) are mixed-motive games. This term draws attention to the complex strategic properties that motivate the players partly to cooperate and partly to compete with one another. A player in a mixed-motive game has to contend with an *intra*personal, psychological conflict arising from this clash of motives in addition to the *inter*personal conflict that exists in the game (p. 93).

In general the *Disarmament Game* may be characterized as a two-person, two-strategy, non-zero-sum, and non-cooperative strategic interaction. It is a sort of game where both parties can simultaneously win or lose to some extent, and in a single play of the game a lose-lose outcome occurs when a win-win outcome might have been possible.

The Disarmament Game, displayed in Table 2.6 below, is played between two countries, Country I and Country II, each of which has two strategies, Disarm and Arm.

TABLE 2.6: DISARMAMENT GAME

(A Mixed-motive Game)

		Country II	
Country I	Strategy	Disarm (Cooperate)	Arm (Defect)
	Disarm (Cooperate)	2, 2	0, 3
	Arm (Defect)	3, 0	1, 1

As Table 2.6 shows, in the *Disarmament Game* the strategy combination (Arm, Arm) yielding the outcome (1, 1) is a *dominant strategy Nash equilibrium* by virtue of the fact that the two players' strategies 'Arm' strictly dominates their respective strategies 'Disarm'. By applying the definition of PSNE, we find that the same outcome (1, 1) is also a PSNE, but none of the three other outcomes is a PSNE. As a result, the outcome (1, 1) happens to be a *unique* PSNE. Yet, this equilibrium outcome is *Pareto-suboptimal* in the weak as well as the strong sense as at least one, and in fact both, players can be made better off without making any one else worse off by a shift from (1, 1) to (2, 2). But the outcome (2, 2), though Pareto-efficient, is not an equilibrium. Thus, the main problem of this game is that there is an *incompatibility* between two extremely intuitively reasonable principles, viz., the *dominance principle* of individual rationality that prescribes the

outcome (1, 1) and the *Pareto Principle* of collective rationality that prescribes (2, 2) which is a fair and mutually more advantageous outcome than the equilibrium outcome (1, 1).

It may be noted that the *Disarmament Game* discussed here can be used to explain the Arms Race that usually occurs between any two rival countries such as, India and Pakistan, Israel and Arab nations, USA and the former USSR, each of which tries to beat the other in the development and accumulation of more and better arms. In fact, this model is just a special application of the more general and famous game known as the *Prisoner's Dilemma* (PD) which has numerous applications in all spheres of life and has been the most widely studied game ever since it was consciously conceived as a general form of game that reflects many real life situations involving elements of both conflict and cooperation. We make a more detailed study of this game in Chapter 4.

Having completed our survey of the most basic types of games that are essential for analyzing the two models that we take up in Chapter 4, we will now make some *critical observations* on two fundamental conceptual tools of game theory, viz., the *Nash equilibrium* and the *Pareto-optimality Principle*. The *Nash equilibrium* concept takes account of the *competitive aspect* of strategic interactions and is basically concerned with finding out the best possible reply of each player given the strategy choices of all others so that the players can all settle for some possible outcome. On the other hand, the *Pareto Principle* which allows people to own as much wealth as they can without lowering anybody else's lot is obviously grounded on the laissez-faire or free market capitalism that recognizes maximum freedom for business enterprises as well as consumers' sovereignty.

Though the Pareto Principle is widely used in economics and engineering, it has certain limitations as a normative criterion for judging alternative allocations and policies. First, *PSNE* does *not always exist*. As we have seen earlier, the game of Matching Pennies, for example, does not have *PSNE*. Second, *MSNE* does *not always exist*. For example, there is no MSNE for the Disarmament Game just because each player has a dominant strategy that is played 100% of the times leading to a dominant strategy equilibrium. Third, Nash equilibrium is *not always socially optimal*. For example, the only Nash equilibrium outcome in the Disarmament Game is Pareto-suboptimal and mutually disadvantageous to the players when compared to another outcome that is not an equilibrium but yet a Pareto-superior outcome. Fourth, in some games, such as the Driving Game and the Stag-Hunt Game that were considered earlier, the number of Nash equilibrium outcome is *not unique* but multiple. Fifth, the existence of *multiple Nash equilibria* of which one is *Pareto-superior* and *payoff-dominant* when compared to the other, as in case of the Stag-Hunt game, and yet they may lead to *coordination failure*. Sixth, Nash equilibria are *not always fair*. For example, the game of Chicken to be discussed in Chapter 4 has two Nash Equilibria which are not equally liked by the two players.

Seventhly, a new study has shown that even if a Nash equilibrium theoretically exists, it may remain *unattainable* as it is often impossible for game players to identify and reach it efficiently. In most cases calculating the Nash equilibrium is a complicated procedure. It is difficult to see how players reach the equilibrium especially in a single play of games. As Klarreich (2017) puts it:

> When players are at equilibrium, no one has a reason to stray. But how do players get to equilibrium in the first place? In contrast with, say, a ball rolling downhill and coming to rest in a valley, there is no obvious force guiding game players toward a Nash equilibrium (*Quanta Magazine,* (July 18, 2017).

Eighthly, the concept of Nash equilibrium has the built-in concept of an instrumentally rational person seeking to maximize self-interest, and so is devoid of any ethical content.

We will now briefly discuss the *Pareto-optimality principle* which is a fundamental concept in welfare economics and is mainly concerned with evaluating alternative economic policies or resource allocations. Though the Pareto principle is extremely popular among economists, it has several serious limitations. We briefly point out some of the most important criticisms of this principle.

Firstly, in some contexts several different alternative allocations may satisfy the definition of Pareto-efficiency, but the Pareto Principle by itself cannot help us to choose between any two possible Pareto-efficient distributions. Thus the Pareto Principle itself cannot tell us about how to choose between the two Pareto-efficient outcomes (Right, Right) and (Left and Left) in the Driving game which are equally good for both players. Similarly, in the game of Chicken, as will be seen in Chapter 4, there are two different Nash Equilibria which are not equally liked by the two players but the Pareto Principle itself cannot judge them as better or worse.

Secondly, there is no guarantee that Pareto-efficient distributions will necessarily be *equitable* or socially desirable allocations, because Pareto efficiency and equity are

unrelated concepts. For example, an allocation that assigns all the goods in the world to me is Pareto-efficient, since there can be no redistribution that can make someone better off without making me worse off.

Thirdly, although the Pareto Principle sounds quite persuasive, its usefulness as a theory is extremely limited as it fails to make any *interpersonal comparison* of utility. As Baumol (1982) points out, the Pareto criterion cannot be applied to judge any policy proposal which will help some but harm others.

A Nash equilibrium outcome is desirable because it allows the players to find a stable position where the various competing forces are resolved. And a Pareto-efficient outcome is desirable, because it marks an improvement in the overall situation without doing any harm to anybody. But Nash equilibrium outcomes and Pareto-efficient outcomes often do not match. For example, in the Disarmament Game the outcome (2, 2) is Pareto-efficient but not Nash equilibrium, while the outcome (1, 1) is Pareto-inefficient but not Nash equilibrium.

Thus neither Nash equilibrium nor Pareto-efficiency is designed to guarantee or even to take account of *distributive justice* of any outcome or allocation of resources. Nash equilibrium is concerned only with where rationally self-interested individuals can settle, while Pareto Principle is a criterion for judging efficiency and is concerned only with raising the wealth level of some and not at all with improving the lots of the least fortunate members of society. To solve this or any other dispute regarding efficiency vs. equity of alternative allocations we must ultimately take recourse to moral philosophy.

2.4.2 Social Dilemmas

This subsection introduces the concept of *social dilemma* and its *classification*, and then discusses its two main types, viz., the *public goods dilemma* and the *commons dilemma*.

As against the classical economic thought that the pursuit of self-interests by the members of a society is not merely consistent with but will rather be transformed into the maximization of the interests of society as a whole by the *invisible hand* of the free market (Smith, 1776), many scholars such as, Arrow (1953), Olson (1965), Hardin (1968), and Sen (1970) convincingly argue that what is individually rational to pursue often clashes with what is collectively rational to pursue. The inconsistency between the individually rational pursuit of immediate self-interests and the long-term collective interests of all the members of society is known as 'social dilemma" also called the "problem of collective action", the "social trap", the "public goods game", and the "tragedy of the commons".

Early philosophers such as, Hobbes (1651) and Hume (1740/1888) clearly conceived of what we now call the collective action problem or social dilemmas. Dawes (1980) defined social dilemma as follows:

> Social dilemmas are characterized by two properties: (a) the social payoff to each individual for defecting behavior is higher than the payoff for cooperative behavior, regardless of what the other society members do, yet (b) all individuals in the society receive a lower payoff if all defect than if all cooperate (p. 170).

Dawes' definition of social dilemma states that all players have a dominant strategy and the pursuit of self-interests by all results in a deficient equilibrium where everybody becomes worse off than when they all cooperate. This definition thus restricts social dilemma to the Prisoner's Dilemma game. But in fact there are many other types of social dilemma such as, the Stag-Hunt game and the game of Chicken. We, therefore, need to broaden the definition in order to take account of these and other types of social dilemma.

A broader definition of social dilemma has been offered by a number of authors (Liebrand, 1983; Kollock, 1998; Shankar & Pavitt 2002; Capraro, 2013; Raub, Buskens, & Corten, 2015). By relaxing the dominance assumption, a *social dilemma* may be defined as a situation where the pursuit of immediate self-interests by the individually rational and strategically interdependent members of a society is inconsistent with the long-term collective interests and leads to an outcome where everyone is worse off than had they cooperated. To be more concise, a social dilemma is a situation where the individually reasonable behavior leads to an outcome in which everyone is worse off.

Liebrand (1983) fomulated the following definition of social dilemma that is broad, and yet precise and easy to apply in various game theoretic models to test if there is any social dilemma involving a conflict:

> … a social dilemma is defined as a situation in which (1) there is a strategy that yields the person the best payoff in at least one configuration of strategy choices and that has a negative impact on the interests of the other persons involved, and (2) the choice of that particular strategy by all persons results in a deficient outcome (p. 124).

The payoff matrices of the Disarmament Game (i.e., PD), the Stag-Hunt, and the Chicken satisfy our definition as well as the two conditions stated in Liebrand's definition and so, qualify as social dilemma. Let us see why the Stag-Hunt game, for example, that has no dominant strategy equilibrium also involves a social dilemma by the broad definition that we formulated. For both players it is individually reasonable to choose the safe strategy Hare which when chosen by both lead to the outcome (1, 1) in which everyone is worse off than in (3, 3) that could have been achieved had they both chosen Stag.

Many of the individual, societal, or global problems that we face nowadays are basically social dilemmas. The widespread existence of social dilemmas has been concisely stated by Kollock (1998) as follows: "Many of the most challenging problems we face, from the interpersonal to the international, are at their core social dilemmas" (p. 183). Dawes (1980) identifies three basic types of social dilemmas each of which has different instances:

> Some of these examples come from the three crucial problems of the modem world: resource depletion, pollution, and overpopulation. In most societies, it is to each individual's advantage to use as much energy, to pollute as much, and to have as many children as possible (p. 171).

Thus, for example, the general problem of *resource depletion* might appear as the specific problems of overfishing, deforestation, depletion of fossil fuels and minerals, soil erosion, and overconsumption of resources.

To figure out the anatomy of social dilemmas, it is important to realize that they are basically mixed-motive games. Hence, social dilemmas may be divided into two broad classes: (1) *Two-person* (or, *Dyadic*) *Social Dilemmas*, and (2) *Multi-person* (or, *N*-

person) *Social Dilemmas*. There are various forms of Two-person Social Dilemmas involving a conflict between two persons' choice of strategy, but the key ones are the Prisoner's Dilemma, the Stag-Hunt (or, Assurance) Dilemma, and the Hawk-Dove (or, Chicken) Dilemma. But the most important social dilemmas that have national as well as international implications are the Multi-person Social Dilemmas which are the multi-person versions of the basic two-person games such as, the Prisoner's Dilemma, the Stag-Hunt, and the Chicken. In fact, the bigger the size of a social group in terms of the number of its members, the greater the conflict between individual interests and collective interests. The two multi-person social dilemmas that we will now discuss are known as the *Public Goods Dilemma*, and the *Common Goods Dilemma* (or, the *Tragedy of the Commons*).

But to discuss the Public Goods Dilemma and the Common Goods Dilemma it is essential to explain the four-fold classification of economic goods into Private Goods, Congestible Goods, Club Goods, and Public Goods on the basis of the presence or absence of each of these two properties of goods, viz., *rivalry* in consumption and *excludability*.

A good or service is ***rival*** in consumption if and only if one person's use of a particular unit of it 'uses it up' so that no one else can simultaneously consume it. For example, an *orange* is rival in consumption, because one person's consumption of it precludes another person from consuming the same orange. On the other hand, a park is nonrival in consumption, because one person's enjoyment of it does not use it up and so does not diminish or destroy another person's ability to consume it. A good is ***excludable*** if and only if it is feasible to prevent a person or group of persons, usually those do not

pay for it, from using it. For example, a cinema is an excludable good if those who do not have a ticket are prevented from entrance. A street light, on the other hand, is not excludable, because there is no way to make the light shine on some but not others walking along the street at night.

Thus *rivalry* is defined from the perspective of the consumer, while *excludability* is defined from that of the producer. Having defined these two different properties of goods, we can now define the four categories of goods.

A ***Private Good*** is one that is both *excludable* from non-payers and *rival* in consumption. For example, an ice cream is a private good, because it is both excludable and rival. It is excludable as one who does not pay for it can be prevented from having it. It is also rival in consumption, because if one person eats it, another person cannot eat the same one. Most goods we consume in society are private goods. The *problems* of scarcity or abundance of private goods are said to be efficiently and effectively *solved* through the optimal allocations of those goods by a well-functioning market system. One problem is that essential private goods are not always available to the poor because of high prices. Another problem is that the production and consumption of private goods such as, cars and many other luxury goods, may be socially suboptimal due to the negative spillover effects on the third party who are neither the producers nor the consumers of those goods.

A ***Public Good*** is the opposite of private good and is neither *excludable* from non-payers nor *rival* in consumption. For example, *national defense* is a public good, because it is neither excludable nor rival. It is non-excludable as one cannot be prevented from having the protection of national defense regardless of whether one does or does not pay

for it. It is also non-rival in consumption, because one person's having the protection of national defense does not preclude another person from having the same benefit. To see what an extremely important role public goods play in our lives, it is enough to ponder over the fact that we cannot drive our car, a private good, to go from one place to another without the road which is a public good.

The non-excludability property of public goods creates the problem of *free riding*. If you create a public good, such as, building a connecting road from your home to the main road, then it will produce positive spillover effects, or to use a technical term positive externality, on others who would benefit from the road but would still not share the cost because of non-excludability. This phenomenon of enjoying the benefits of some good without bearing burden of producing them is known as the problem of ***free riding***. The traditional solution to this problem is for the state to impose taxation for funding the production of public goods. Though taxation attempts to solve the problem of free riding and provides the public good, the question of how efficiently and how equitably the government performs these functions remains an important field of investigation.

A ***Congestible Good*** (or, ***Common Good***) is one that is rival in consumption but not excludable from non-payers. For example, fish in the sea are congestible or common goods, because they are rival in consumption but not excludable. They are rivalrous in consumption, because when one person catches fish, another person's ability to catch or consume fish is reduced. But given the vast size of the sea, they are not excludable. Another example of a congestible, i.e., non-excludable but rival, good is a common pasture where people graze their cattle.

A ***Club Good*** (or, *Artificially Scarce Good*, or *Natural Monopoly*) is one that is excludable from non-payers but not rival in consumption. Cable TV, private parks, and computer software are examples of club good. A Cable TV, for example, is a club good, because it is a good for which it is feasible to prevent those who have not paid for it from having access to it, but whose consumption by one person does not prevent the simultaneous consumption of it by others. A club good being non-rival in consumption can be provided to additional consumes at zero marginal cost, but yet non-payers are prevented from having access to it just because doing so is possible and profitable for the producer. That is why a club good is also called an "artificially scarce good" or a "natural monopoly". The problem associated with club goods is called the "Tragedy of Artificial Scarcity" which is a case of *market failure* where a market by itself cannot produce the socially optimal outcome.

The division of goods into the four categories discussed above is shown in Table 2.7 below. However, the borderlines between these classes may often appear to be vague rather than clear-cut. The reason for this fuzziness of the classes is that they are defined in terms of the two properties of rivalry and excludability which are not absolute but rather a matter of degree. In fact, we can think of two continuous spectrums, one for excludability and another for rivalry, along each of which the values of one variable vary from the highest to the lowest. Thus we can define a ***quasi-public good*** (or, a near-public good) as one that is partially rival and partially excludable. A *quasi-public* good thus shares the attributes of both a private good and a public good. For example, a road which at normal times works as a *public good* may at peak hours turn out to be a *quasi-public* good.

TABLE 2.7: FOUR TYPES OF GOODS – PROBLEMS & SOLUTIONS

Product Property	**EXCLUDABLE** from non-payers	**NONEXCLUDABLE** from non-payers
RIVAL in Consumption	**<u>PIVATE GOODS</u>** (or, Market Goods) 1. RIVAL 2. EXCLUDABLE EXAMPLES: Food, Clothing, Toothbrush, Cars, Private Health Care (Merit Good), Education (Merit Good), Congested Toll Roads PROBLEM: How to make a Socially Optimal Allocation SOLUTION: Allocation by Well-functioning Mkt.	**<u>CONGESTIBLE GOODS</u>** (or, Commons/ Open Access Goods) 1. RIVAL 2. NON-EXCLUDABLE EXAMPLES: Fish in the sea (Overfishing), Envir (Pollution), Virgin forest (Deforestation), Common Pastures (Overgrazing), Congested Nontoll Roads (Traffic Jam) PROBLEM: **<u>Commons Dilemma</u>** or <u>Trag. of Commons</u> (PDG, Mkt. failure) SOLUTION: Collective Mgmt. by Insiders without Govt. or Pvt. Control (Elinor Ostrom)
NONRIVAL in Consumption	**<u>CLUB GOODS</u>** (Artificially Scarce Goods/ Natural Monopolies) 1. NON-RIVAL 2. EXCLUDABLE	**<u>PUBLIC GOODS</u>** 1. NON-RIVAL 2. NON-EXCLUDABLE

	EXAMPLES: Cinemas, Cable TV, Private Parks, Computer Software, Patented Medicines, Knowledge in Heads, Wifi, Uncongested Toll Roads PROBLEM: Trag. of Artificial Scarcity (Mkt. failure) SOLUTION: Designate Property Rights & Use cap-auction trade to allocate these	EXAMPLES: Street Lights, Tornado Siren, Lighthouse, Broadcast TV, National Defense, the Law, Uncongested Nontoll Roads PROBLEM: **<u>Public Good Dilemma</u>** – Free Riders' Problem, & Missing Mkt., PDG) SOLUTION: Collect Depletion & Pollution Taxes so that Govt. can Provide these goods.

Mankiw (2018) have given examples to show that what we normally call the same good such as, a particular *road*, may be characterized as one of the four types of good depending on the degree of excludability from non-payers and the degree of rivalry in consumption. Thus, an uncongested non-toll road, i.e., a road that is non-rival in consumption and non-excludable from non-payers, is a *public good*. On the other hand, a congested toll road, i.e., a road that is rival in consumption and excludable from non-payers, is a *private good*. In between these two extremes, there are two other cases. One is a congested nontoll road which is rival in consumption but excludable from non-payers, and hence a *common* or *congested good*. The other one is an uncongested toll road which is non-rival in consumption but excludable from non-payers, and hence a *club good*.

Having explained the four types of goods, we now discuss the ***Common Goods Dilemma*** as a multi-person social dilemma. We have seen that common goods are both non-excludable and rival in consumption. These two properties of a common good together lead to an excessive use and eventual depletion of the good which has come to be known as the "tragedy of the commons". The consumers of a common good, such as a pasture, will continue using it more and more as long as it provides them with some positive benefit just because it is non-excludable. But the consumers of the common good will also continue getting less and less marginal benefit or utility from it just because it is rival in consumption Thus the continued use of the common good which is non-excludable and rival in consumption will lead to an overuse and ultimately to its depletion unless appropriate and timely steps are taken for their proper management.

Perloff (2000) pointed out two possible solutions to the *multi-person common goods dilemma* also known as the *tragedy of the commons*. One possible approach is direct control by the government through passing law that requires either the users to pay tax or fees as compensation for the harm they do to the common good or restricting or putting temporary ban on access to the good. The second solution advocated by Hardin (1968) consists of breaking up a large common pool resource into smaller and independent units and then assigning private property rights to each unit. This solution is like transforming a common good into a private good. But the third and best possible solution for the problem of common goods has been offered by Elinor Ostrom, a political scientist, who happened to be the first woman to receive the Nobel Prize in economics. Ostrom (1990) showed that common goods can be better managed not by the outsiders or government agencies but rather collectively by the insiders who are physically close to

the goods, related to each other, more aware of the local conditions or norms, and well-equipped to self-police to guarantee that all members obey the rules of the community.

We will finally deal with the second of the two most important multi-person social dilemmas known as the ***Public Goods Dilemma*** (or, *Public Goods Game*) that can serve as a model for many real-life situations involving conflict between the reality of self-interest and the need for cooperation for funding public goods like public hospitals, public libraries, bridges, public television stations, clean environment, and so on. Since a public good, as noted earlier, is non-rival and non-excludable, all members of a society can benefit from it but no one would be willing to contribute for it. The presence of greed for superseding others, on the one hand, and the fear of becoming a sucker provide each person with a reason for free riding on the contribution of others.

Numerous authors (for example, Carter, 2007; Irwin, 2009; Ale, Brown, & Sullivan, 2018) have illustrated the concept of the Public Goods Game with essentially the same but superb numerical example. Let us assume that there is a group of *four persons* each of whom is offered *$10* and told that each can freely allocate the money between a private account and a group investment project and that any money kept in private account will remain unchanged while the amount of total money invested will be *doubled* and *equally divided* among all of them *regardless of who invested how much*. This implies that the investment made by anyone will generate a positive externality on every other member of the group including even those who invested nothing.

Now, the *socially optimum outcome*, i.e., the outcome that yields the maximum total payoffs for all, can be achieved if everyone invests the entire initial amount of $10.

This would, according to the rules of the game, yield $20 for each person. But the *socially least desirable* outcome would happen only if no one invests any amount of money which would leave everyone with the initial amount of $10. It is important to point out that the socially least desirable outcome happening when none contributes any money at all is a ***Nash equilibrium*** because none would like to unilaterally switch to any other outcome. However, things would be totally different if looked at from an individual's perspective. For example, the worst outcome for an individual takes place not when nobody invests any amount of money but rather when she invests all her $10 and everyone else invests $0 which would yield a net loss of $5 for the sole investor and a net gain of $2½ for each of the other three. Similarly, the best as well Pareto-optimal outcome for any one individual would take place only when all the three others invest their entire amount of money and she invests nothing but free rides on others' contributions which would leave her with a total of $25 including a net gain of $15. This conflict between collective rationality that demands contribution to the public good and individual rationality that requires free riding on other's contributions gives rise to the ***Public Goods Game*** that has been widely studied by social and biological scientists and is now being studied even by engineers.

Now, a lot of different questions could be raised about how individuals as members of different types of group would behave when facing a public goods dilemma. According to the traditional economic models of *homo economicus*, rational individuals would always choose to free ride. A question may be raised as to whether people are basically egoistic or selfish by nature. Even if we assume that people are by nature egoists, it may be asked whether it is selfishness that causes people to behave non-

cooperatively. Now, if we take for granted that people's egoistic nature is the cause of non-cooperative behavior, it may be asked, as pointed out earlier, whether altruism is a necessary condition, a sufficient condition, both necessary and sufficient condition, or neither. If we assume or can find out that altruism is neither a necessary nor a sufficient condition for cooperation, we could still ask whether altruism is altogether irrelevant or somehow acts as a contributing cause of cooperation. Axelrod (1984) has shown that even purely egoistic players will cooperate if they have the opportunity to meet repeatedly for an indefinite number of times. There are, however, cases when strangers cooperate with each other even if there seems to be no chances that their paths will ever cross in the future.

2.5 Natural Selection and the Darwinian Puzzle

This section is intended to be a brief one that tries to explain and evaluate Darwin's theory of natural selection and the Darwinian puzzle that actually initiated the search for the mechanisms of cooperation and thus led to the development of the various theories of cooperation. In subsection 2.5.1 we present a short account of Malthus's theory of population as a background to Darwin's thought. Then in subsection 2.5.2 we explain Darwin's theory of natural selection and in subsection 2.5.3 we present a brief criticism of his theory.

2.5.1 Malthus as Precursor of Darwin

In his influential book *"An Essay on the Principle of Population, as it Affects the Future"* the economist Thomas Robert Malthus (1798) painted a grim picture of human society characterized by selfish behavior of individuals and a relentless struggle for survival. Malthus claimed that the population of a country is limited by the availability of food and that population increases faster than food, because human *population* increases every twenty-five years in a geometrical ratio, such as 1, 2, 4, 8, 16, ... and so on, while *food* supply increases in an arithmetical ratio, such as 1, 2, 3, 4, 5, ... and so on. So, in the course of time population outruns food supply. Hence, he concluded that unless humans apply *preventive checks*, such as late marriage, self-restraint, and celibacy, to lower the birth rate, nature will activate *positive checks*, such as extreme poverty, tormenting occupations, bad nursing of children, epidemic, war, and famine, to raise the death rate for limiting the growth of population. Thus Malthus's central theme is that the tendency of humans to produce more offspring than the carrying capacity of the means of subsistence available creates a perpetual state of hunger, disease, and struggle. Thus it is obvious that Malthus (1798, Ch.10), like the other classical economists, views humans as being essentially selfish and maintains that in a state of constant struggle and a search for the means of survival it is natural that selfishness would become the dominant strategy.

2.5.2 Darwin's Theory of Natural Selection

The theory of evolution by natural selection developed by Darwin (1859) deems ruthless competition to be a common phenomenon and requires individuals to be

competitive rather than cooperative in order to win in the struggle for existence by gaining control over the limited means of survival. As Darwin describes:

> … as more individuals are produced than can possibly survive, there must in every case be a struggle for existence, either one individual with another of the same species, or with the individuals of distinct species, or with the physical conditions of life. It is the doctrine of Malthus applied with manifold force to the whole animal and vegetable kingdoms; for in this case there can be no artificial increase of food, and no prudential restraint from marriage. (Darwin 1859:63).

Darwin's theory was based on a developing trend of thought that questioned the previous concepts of the natural world and essentially changed the course of future scientific thinking. Darwin read Malthus's theory of population and was deeply influenced by the crucial concept of constant struggle for survival out of which he developed his theory of evolution through natural selection. He presented his theory with convincing evidence in his book entitled "*On the Origin of Species By Means of Natural Selection* (Darwin, 1866). The process of *natural selection* may be better understood by contrasting it with the somewhat similar process of *artificial selection* which is selective breeding of organisms carried out under controlled conditions by humans for the development and perpetuation of desirable traits in subsequent generations. Ridley (p. 682.) highlights the fact that the "forms of most domesticated and agricultural *species* have been produced by artificial selection". The main point of difference between the two processes is that in *natural selection* nature causes the changes in the species, while in *artificial selection* humans play an active role by making a deliberate attempt at bringing about some desirable changes in the species.

The essence of Darwin's theory of evolution is a mechanism called natural selection by which populations of organisms with variations in traits that better enable them to *adapt* to their environments live longer, compete better for food or mates, and reproduce more offspring than populations that do not have the variations, thus ensuring the perpetuation of those favorable traits in succeeding generations.[10] To take a simple example of evolution through natural selection, consider an ecosystem where there are birds that feed on bugs. Suppose there red bugs and green bugs, and the birds prefer red bugs to green bugs just because red bugs are tastier or easier to detect in the environment. So, the percentage of red bugs will decrease and that of green bugs will increase. The green bugs will reproduce and multiply. Therefore, red bugs will eventually be wiped out and only green bugs will be left.

The theory of natural selection can be regarded as an *argument* where the conclusion follows logically from a set of premises as explained below[11]:

1. Reproduction. Organisms within a species produce offspring and thus create a new generation.
2. Heredity: The offspring tend to resemble their parents. That is, like is likely to produce the like. But only some traits are consistently passed on from parents to offspring. Such traits are inheritable, whereas other traits are strongly influenced by environmental conditions and show weak inheritability.

[10] Besides natural selection, there are three other basic mechanisms of evolution such as mutation, migration, and genetic drift. For a lucid but brief explanation of these four mechanisms of evolution see *Understanding Evolution*, 2016.

[11] The conditions are stated slightly differently by different authors. See, for example, (Ridley, p.74) and (Mikkelsen & Robin, p.104).

3. Variation: Although offspring resemble their parents, careful observation shows that no two individuals of a species, except perhaps identical twins, are exactly the same with respect to external features such as appearance and behavior. That variation in traits arises out of new combinations of genes that occur when organisms reproduce sexually. These variations may occur in body size, hair color, voice properties, or the number of offspring, behaviour, and so on. On the other hand, some traits, such as the number of eyes in vertebrates, or sex show little or no variation among individuals.
4. Overpopulation: Organisms of most of the populations have high fecundity and so produce more offspring each year than can survive in the given environment.
5. Scarcity: The quantity of resources such as food, habitat, or mates available for the sustenance of the entire population of any species is inadequate.
6. Struggle for Existence: The struggle for existence refers to environmental competition. The abundance of the members of the population on the one hand and the scarcity of resources on the other lead to a relentless struggle for resources necessary for survival.
7. Variation in Adaptation: Individuals differ in adaptation or fitness. Adaptation, as defined by (Ridley, p. 682), is a "feature of an organism enabling it to survive and reproduce in its natural environment better than if it lacked the feature." Thus individuals inheriting those variable traits which favor them in the struggle for local resources are better *fitted* to survive and reproduce than others. Species whose individuals are the fittest will survive while others become extinct.

8. Genetic Composition: In the succeeding generations there will be a higher frequency of individuals possessing favorable traits.

Darwin's Conclusion: Thus, according to Darwin, evolution by the blind mechanism of natural selection occurs only when certain conditions, such as reproduction, inheritability, variation in traits, and variation in adaptation, are satisfied. If the favorable traits are more likely to pass on from progenitor to progeny, it follows that over time individuals with favorable traits will become more common in the population while those with unfavorable traits will gradually disappear. Repeated occurrence of natural selection to many generations over a long period of time gradually change the population to adapt more and more to the environment, and these small variations accumulate over time and ultimately result in the emergence of a new species.[12] Strong evidences of natural selection have been obtained from observation and fossil record. It may be noted that natural selection operates on comparative rather than absolute advantage. As Darwin (1866) points out, "…as natural selection acts by competition for resources, it adapts the inhabitants of each country only in relation to the degree of perfection of their associates".

2.5.3 Limits of Darwin's Theory

Darwin himself was aware of several problems with his theory of evolution. As stated by Hampton (2009) there are three different problems, viz., the *'problem of non-*

[12] For a simple explanation of the emergence of new species see (Mikkelsen & Robin, 2011).

fitness', the problem of *mechanism of inheritance*, and the problem of *altruism*. The 'problem of non-fitness' relates to the obvious fact that there are species having some typical traits that do not help but hamper their survival. For example, the size and brightness of the male peacock's tail is costly in terms of energy to grow and maintain it and makes it more vulnerable to predators. Darwin (1871) solved this problem with his theory of *sexual selection* which holds that those individuals some of whose traits may inadvertently attract predators can display their fitness to be chosen by interested members of the opposite sex and thereby increase the chances of their mating and reproductive success. As to the problem of the mechanism of inheritance, Darwin did not know how adaptive variations were transmitted from parents to progeny He was also unfamiliar with Gregor J. Mendel's (1822–1884) discovery of the mechanism and rules of inheritance which together with the Darwinian theory of evolution form the theoretical basis of modern biology. The third problem that puzzled Darwin much was the problem of altruism which refers to the apparent lack of selfishness exhibited by many of the species he observed.

Let us take a closer look at Darwin's puzzlement about the problem of altruism. A guiding principle that Darwin gleaned from Malthus was that the population growth for any species is always constrained by resources. It seems to follow that for survival and reproduction in such circumstances organisms must *compete* with one another for the limited resources, such as foods and habitats, which sustain life. Thus the persistent *resource shortage* together with the inevitable *competition* for resources implies that organisms should be extremely *selfish*. In fact, Darwin himself was aware that there is an apparent *contradiction* between the *selfish selection* view of natural selection and the

observation of cooperative and sacrificial behavior of some individuals to benefit others amongst various species such as ants, birds and primates. Now, according to natural selection, if an organism sacrifices the resources it needs for survival and reproduction in order to benefit another organism, then that particular predisposition would, by assumption, not be transmitted to the succeeding generation. Thus the conclusion that necessarily follows is that eventually *selfishness* will outcompete and eliminate *selflessness*.

But Darwin himself is inclined to admit that natural selection may operate also at the level of the family or group when he, as Gayon (1998) points out, broadens the 'individualistic concept of natural selection' and writes: "This difficulty, though appearing insuperable, is lessened, or, as I believe, disappears, when it is remembered that selection may be applied to the family, as well as to the individual, and may thus gain the desired end." (Darwin, 1866, p. 284).

By emphasizing constant struggle, natural selection suggests unconditional defection (All-D) as the most rational and successful strategy for an individual. But in the real world most interactions are *repeated* and take place within *settled communities* of animals or humans where both defectors and cooperators are soon detected and well-known and as a result All-D cannot serve as a successful strategy (King, 2015, pp. 14-17). In a *stable* society defectors soon meet with defections from all others and hence end up with unnecessarily low scores, or in other words, lower prospects of survival and reproduction. Moreover, defectors even run the risk of facing what is called 'social ostracism' when other players refuse to play with them. There is, however, a little

likelihood of success for a defector only when the interaction takes place with a stranger and it is known that their paths will not cross again in the future.

Thus natural selection by itself cannot explain the evolution of cooperation. Though Darwin could not explain the origin of cooperation, he was struck by the co-presence of conflict and cooperation and so may be said to have broadened the concept of natural selection from the individualistic to the group level when he writes, “This difficulty, though appearing insuperable, is lessened, or, as I believe, disappears, when it is remembered that selection may be applied to the family, as well as to the individual, and may thus gain the desired end.” (Darwin, 1859, p. 237).

2.6 Theories of the Mechanism of Cooperation

This section is divided into seven subsections each of which discusses a distinct type of mechanism of cooperation.

2.6.1 Kin Selection

The *kin selection theory* also known as *inclusive fitness* theory deems natural selection to be taking place not at the group level but at the individual or rather genetic level. This theory was anticipated by Haldane when he, in reply to the question if he would give his life to save a drowning brother, told that he would jump into a river and risk his life to save 2 brothers but not 1 or 8 cousins but not 7 (McElreath & Boyd, 2007, p. 82). Although the *inclusive fitness* theory was proposed by Haldane, it was actually

elaborated by Hamilton (1963, 1964) and renamed *kin selection* theory by Maynard Smith (1964). According to Hamilton, natural selection leads to an increase in an individual's inclusive fitness. By *fitness* is meant the reproductive success or the ability of an organism to survive and to produce offspring and thus to propagate its genes to the next generation. Depending on how an individual reproduces it, there can be two different kinds of fitness – direct and indirect. *Direct fitness* of an individual is measured by the number of surviving offspring produced by that individual. *Indirect fitness* of an individual, on the other hand, is measured by the number of offspring contributed to the next generation by close relatives of the individual who shares genes with them and offers them essential help to achieve additional reproductive success. The combination of an individual's direct and indirect fitness is *inclusive fitness*.

Hamilton (1964) puts forward a general formula known as "Hamilton's rule" that captures the costs and benefits of altruism and their relationship to kinship and states that altruism is favoured by natural selection when

rB > C [or, (rB – C) > 0, or, r > C/B]

where

r = Coefficient of genetic relatedness between the altruist and the beneficiary measured by the proportion of genes shared by the two individuals (e.g., ½ for siblings, ¼ for nieces and nephews, ⅛ for cousins and so on)

B = Benefit to the beneficiary of the altruistic act measured by how many more offspring are produced by the beneficiary as a result of help from the altruist

C = Cost to the altruist measured by how many fewer offspring are produced as a result of the altruistic act

Hamilton's rule implies that helping a relative increases the inclusive fitness of the altruistic actor by an amount (rB – C), and that if (rB – C) is positive, then the higher its value the more the helping behavior will increase. Thus kin selection will occur when the measure of genetic relatedness exceeds the cost-to-benefit ratio. In other words, kin selection occurs when an actor's inclusive fitness increases due to an altruistic act that enhances the reproductive success of relatives. Let us take a couple of classic examples of the inclusive fitness view of altruism (Dawkins 2006/2010; "Kin selection", in *Wikipedia*, n.d.). One example is provided by the ***sterile female eusocial insects*** that can acquire reproductive benefits not by themselves reproducing but by helping their relative, the reproductive queen, propagate their shared genes to the future generations. Another example consists in the ***alarm calls*** any member of a group gives to the other members to warn them of approaching predators which involves the risk of attracting the predator's attention to the caller. There are plenty of examples of human altruism through kin selection as humans tend to behave more altruistically to kin than to unrelated individuals.

But a question can be raised whether the kin selection theory is immune from the problem of sub-optimization. Using a variant of the story attributed to Haldane (1955), **Zahavi** (1995, 1) has presented an ingenious argument to expose a fallacy of the kin selection theory of evolution of altruism and cooperation:

> … if one of two brothers walking beside a river, were to fall into it and be in danger of drowning, it would be reasonable for the other brother to risk his life somewhat to save the drowning brother, since by taking such a risk (i.e.

decreasing his fitness), he may save his brother and increase the frequency of genes similar to his own in the following generation.

The instability of the model is clearly apparent if the same story is told with three or more brothers, rather than two, walking along the river. It is obvious that if one of them jumps to the rescue, the other sibling (who does not risk himself), gains as much as the one who risks himself, but without incurring any cost.

Thus the kin selection model cannot offer a satisfactory account for the evolution of cooperation in so far as the total fitness gain of the selfish brother who happens to be a social parasite is higher than that of the altruist brother. Moreover, the kin selection theory cannot explain the evolution of cooperation in general and the cases of costly cooperation among competitors who are genetically unrelated in particular.

2.6.2 Group (or Multilevel) Selection

The Group (or Multilevel) Selection theory holds that natural selection acts not only on individuals but it can also simultaneously act on multiple levels of life such as cells, individuals, or groups (Okasha, (2006). This implies that any behavior that hurts a given individual but helps other individuals might evolve if it is beneficial at a higher level such as the group. The theory of **group selection** which was first strongly advocated by **Wynne-Edwards** (1962) and popularized by **Ardrey** (1970) holds that individual organisms act in such a way as to promote not their personal interests but the survival of the group or species to which they belong and thus implies that natural selection acts via group selection rather than individual selection or what **Dawkins**

(2006/2010) calls 'gene selection'. To see the argument for group selection, assume that there are two groups of individuals such that one group is composed of more cooperative individuals, while the other of more selfish individuals. Now, due to the law of synergy[13] the members of the cooperative group will together achieve more resources than the sum-total of resources that can be individually made. Hence, the group of self-sacrificing cooperators will be better fitted to survive and will eventually outcompete and eliminate the group of selfish non-cooperators.

It may be noted that even Darwin himself recognized the importance of group selection when an individual member's moral character is helpful for the survival of the group:

> "There can be no doubt that a tribe including many members who … were always ready to give aid to each other and to sacrifice themselves for the common good, would be victorious over other tribes; and this would be natural selection" (Darwin 1871, 166).

But the above picture of the optimistic situation drawn by group selection theory overlooks the existence of 'suboptimization' which does not generally lead to global optimization (Machol, 1965, pp. 1-8). Heylighen (1992), Zahavi (1995), and Dawkins (2006/2010 reprint) have clearly argued that there is a subtle conceptual error in the reasoning for group selection as it ignores the incompatibility of 'suboptimization' with global optimization. Due to variability of inherited characteristics there will be

[13] *English Oxford Dictionaries* (https://en.oxforddictionaries.com/definition/synergy) define synergy as "The interaction or cooperation of two or more organizations, substances, or other agents to produce a combined effect greater than the sum of their separate effects." Usually, cooperation can create a synergy which means that the members of a group can together achieve a whole greater than the sum of the simple parts that the members of the group can separately achieve.

differences in 'cooperativity' among the members of the group. Now within the group the relatively selfish members will free-ride on the relatively altruistic members by enjoying the benefits of the latter's sacrifice but without bearing any burden of reciprocal sacrifice. The selfish members will eventually become 'social parasites' and will be fitter than and wipe out the altruists within the group. That is why E.O. Wilson (2014) claims, "Within groups selfish individuals beat altruistic individuals, but groups of altruists beat groups of selfish individuals." Thus the strategy of self-sacrifice would be, to use a concept introduced by Maynard Smith and Price (1973), an *evolutionarily unstable strategy* in the sense that if it is adopted by the population it can be invaded by the alternative strategy of selfishness adopted by a small number of mutant members, just because the egoists take advantage of the altruists' sacrifice but do not give anything in return.

2.6.3 Spatial Selection

Spatial selection is a mechanism for cooperation which is based on the spatial structure of population and can be expressed by the maxim "Neighbors help each other". This means that bunches of cooperators who stick together to form clusters can 'prevail against exploitation' by defectors. But the ability of a cluster of cooperators to win against the adversaries depends on the strength of cooperation within the group.

But the development of successful cooperation depends on the population structure (Nowak and May, 1992). Thus strategies that are successful in a mixed population may not be suitable for a structured population and vice versa. The spatial structure of population specifies the vital acts of social interactions as well as

reproduction. The various ways in which the spatial structure of population affects the evolution of cooperation among the interaction partners form an exciting field of study.

Understanding when and to what degree a spatial structure affects the evolution of cooperation is an important and challenging topic. Su, Li, Wang, & Stanley (2019) have shown that two important factors, viz., *overlapping* role models and *frequency* of interactions among the interaction partners are two very important factors that facilitate cooperation.

2.6.4 Direct Reciprocity

Kin selection theory is obviously unsatisfactory as it cannot explain cooperation between nonrelatives or between members of different species. Such considerations led Trivers (1971) to propose *reciprocal altruism* and Axelrod (1984) to endorse *tit-for-tat* as the most effective strategy for the evolution of cooperation. Both *reciprocal altruism* and *tit-for-tat* are based on the mechanism of ***direct reciprocity*** that embodies the principle 'I help you, you help me'. Direct reciprocity evolves when costs and benefits of cooperation are exchanged directly and repeatedly between any two given individuals. Thus, when A helps B, A bears the cost and B gets the benefit of cooperation, and, in return, when B helps A, B bears the burden and A obtains the benefit.

The reciprocal altruism model designed to explain why non-relatives should make sacrifices to help potential competitors succeed is a form of social interaction whereby one organism acts in a way that temporarily reduces its own fitness by a quantity **c** denoting cost but enhances the fitness of another organism not closely related by a

quantity **b** denoting benefit in expectation of reciprocal treatment in the future. Direct reciprocity can make the evolution of cooperation possible only if the probability, w, of the same two individuals meeting again is greater than the cost-to-benefit ratio of the altruistic act. Formally,

$w > c / b$

where

b = benefit to the beneficiary

c = cost to the cooperator

$b > c > 0$

w = probability that the same players will play again

An example of reciprocal altruism is *cleaning symbiosis* which is a mutually beneficial agreement between a cleaner fish and a relatively large client fish where the former removes and eats parasites from the surface of the latter's body. This is a kind of association that offers opportunities for indefinitely repeated interactions for mutual benefits rather than a one-shot unfair transaction (**Feder**, 1996). Here both the fish benefit – the large client fish gets cleaned by not eating the cleaner fish while the cleaner fish gets food and security by doing the cleaning.

The direct reciprocity model has two important practical flaws. First, it relies too heavily on memory. A reciprocal altruist must remember all the interactions with all the opponents in the past however remote and must be able to recognize all the opponents with whom he or she ever had transactions in the society however large. This is

unrealistic. Secondly, many one-shot encounters happen once in a lifetime between two strangers whose paths are never likely cross again. Direct reciprocity cannot take account of nice or ultra-social behavior in such situations.

2.6.5 Indirect Reciprocity

Indirect reciprocity is a sophisticated and extremely important mechanism for cooperation developed by modern society which, unlike direct reciprocity, does not require repeated encounters between the same two persons. While direct reciprocity is captured by the maxim 'I help you and you help me', indirect reciprocity is based on the dictum 'I help you and somebody else helps me'. Thus indirect reciprocity can explain cooperation between individuals who have never met before and will probably never meet again. Then the question arises as to what the basis is on which the transaction takes place. It is ***reputation*** or social status of an individual that qualifies a person to receive help or cooperation. Reputation is determined by a person's behavior or performance in the past. It reduces the problem of cheating and free-riding in indirect reciprocity. Thus, reputation is a decisive factor in indirect reciprocity.

If society did not develop the mechanism indirect reciprocity, life would have been much poorer than not it really is. However, the need for assessing reputation in indirect reciprocity raises a difficult problem. Determining reputation accurately normally demands high cognitive abilities.

2.6.6 Strong Reciprocity

Strong reciprocity is the inclination of an agent to sacrifice valuable resources of his or her own for rewarding a cooperative or fair behavior and punishing a non-cooperative or unfair behavior regardless of any material rewards for the act of sacrifice (Fehr, Fischbacher, Gächter, 2002). The punishment involved in strong reciprocity may be either second party punishment or third party punishment. In case of the second party punishment, the person who is harmed by the other party's failure to cooperate punishes the non-cooperator. In case of the third party punishment, an uninvolved third party punishes the non-cooperator. The major argument in favor of strong reciprocity is that it works effectively in times of emergency or crisis, such as famine, when neither direct reciprocity nor indirect reciprocity can help for maintaining cooperation.

2.6.7 Costly Signaling

The theory of costly signaling, pioneered by biologist Zahavi (1975), is an informational approach to the study of how mutually advantageous cooperation between two parties may arise through the communication of information from the more informed to the less informed party. Economists have shown that the existence of asymmetric information between buyers and sellers leads to inefficient performance of the economy which can be overcome through the transfer of information by costly signaling (See, for example, Varian, 1992, 2010; Mankiwi, 2018).

Generally, a signal refers to a gesture, action, or sound used to communicate information or instructions from one party to another. For example, traffic lights and

vehicle lights are used to signal information to drivers and passers-by. Firms use advertisements to signal information about the quality of their products to the consumers. It should, however, be noted that the lights or advertisements do not themselves contain any information but only signal information that may be scanned by the receiving party to obtain information.

Now, for a gesture, action, or sound to be an ***effective*** signal, it must fulfill two conditions. First, the signal must be *costly* enough so that the owners of a bad product cannot afford to bear the costs. If a signal were free, then anybody could use it, and it would pass on no information. Second, the signal must be *not too costly* to be borne by the owners of a good product.

Suppose there are two sellers of second hand car. One sells good second hand cars and the other sells bad ones. But the potential buyers of second hand cars obviously have much less information about those cars than the sellers. Now, the buyers would seek information about the quality of those cars, while the sellers would start sending strategic signals about their quality. Thus the sellers of good cars, for example, could offer warranty for satisfactory service of their cars for a certain period of time, because they know that their cars are good. But the sellers of bad cars could not offer any such warranty, because they know that the quality of their product is poor and so the cost of such warranty would be so high that they could not afford to bear it. Thus the mechanism of costly signaling which obviously need not be kept confined to business phenomena provides a sophisticated theory for explaining the emergence of cooperation.

2.7. The Egoism-Altruism Debate

Before going on to examine the Egoism-Altruism Debate, it is important to indicate that the terms "egoism" and "altruism" refer to two opposing theories and are defined in different ways and from different perspectives. They may be defined either as *descriptive* or as *normative* theories of human and animal behavior. They may also be defined either in terms of the *motive* that brings about the act or in terms of the act's *consequences* on the actor (or, agent) and the target of the action (or patient). While egoism is rooted in the Latin word '*ego*' meaning 'I' and emphasizes the impact of an act on the actor's own self, altruism is rooted in the Latin word '*alter*' meaning 'other' and emphasizes the impact of an act on the others. Let us begin by defining the four terms – psychological egoism, ethical egoism, psychological altruism, and ethical altruism – in general language as usually found in standard dictionaries such as, the *Collins English Dictionary*.

Psychological egoism may be understood as the general descriptive statement that everyone has a 'concern for one's own interests and welfare'. *Ethical egoism* is 'the theory that the pursuit of one's own welfare is the highest good'. *Psychological altruism* is the general factual claim that people's behaviors demonstrate 'the principle or practice of unselfish concern for the welfare of others'. *Ethical altruism* is 'the philosophical doctrine that right action is that which produces the greatest benefit to others'.

Neusner and Chilton (2005) point out that the usual dictionary definition of altruism, as opposed to egoism, as "unselfish concern for the welfare of others" has four

components. First, the word "unselfish" implies that an altruistic agent acts for the sake of the others rather than for herself. Second, the word "concern" indicates that altruism entails a "motive" as well as an "act" to which the motive is directed. Third, the word "welfare" means that the motive of the act is to help, rather than harm, someone. Fourthly, the word "others" implies that the recipient of the help is not the actor herself but others.

The characterization of altruism as entailing a motive for being concerned with the welfare of others requires us to confront an important distinction between 'behavioural altruism' (or 'evolutionary altruism') and 'psychological altruism'. Behavioral altruism is defined in terms of the act's consequences on individual fitness or well-being of the helper and the recipient of help, whereas psychological altruism is defined in terms of the internal motives that led to the helping behavior. Thus Hamilton (1964), and his followers working on evolutionary biology define behavioral altruism as an act that raises the fitness of the recipient of altruistic behavior at a cost, i.e., loss of fitness, to the actor. On the other hand, scholars such as Batson & Shaw (1991), Wilson (1992), Clavien (2012), Okasha (2013), emphasize the importance of defining psychological altruism in terms of the motives of the actor. Thus Batson & Shaw (1991) define both altruism and egoism in terms of motivation as follows: "Altruism is a motivational state with the ultimate goal of increasing another's welfare. Egoism is a motivational state with the ultimate goal of increasing one's own welfare" (Batson & Shaw, 1991, p. 108).

Thuss altruism and egoism are similar in two respects. First, both altruism and egoism are goal-directed behavior. Second, both aim at increasing the wellbeing of

someone or others. But they differ in so far as egoism is concerned with someone's own wellbeing while altruism is concerned with the wellbeing of others.

We have pointed out the distinction between two different types of altruism, viz., *biological altruism* which is based on the reproductive or fitness consequences of behavior and *psychological altruism* which is based on motivating intentions of behavior. It is important to note that for explaining human cooperative behavior in philosophy and the social sciences psychological altruism is a richer and more appropriate concept, because our everyday use of the term altruism is more in accord with the concept of psychological altruism than with that of biological altruism.

Let us now take up the Egoism-Altruism Debate which is concerned either with the fundamental descriptive question as to whether man is by nature egoistic or altruistic, or with the fundamental normative question as to whether man ought to behave egoistically or altruistically. In other words, the debate is may be concerned with either one of these two issues. It may deal with the empirical or descriptive issue as to whether or not people *are as a matter of fact motivated* solely by a concern for maximizing their self-interests. Or, it may deal with the normative or prescriptive issue as to whether or not people *ought to be motivated* solely by a concern for maximizing the wellbeing of others without having any ulterior motive for promoting their self-interests. It is often believed that although people primarily attempt at promoting personal proceeds, they can and ought to at times act to fulfill their obligations to other people's rights, privileges, and welfares as required by morality.

But in the famous thought-experiment known as the "Ring of Gyges" and expounded in Book II of the *Republic*, Plato's elder brother Glaucon recalls the legend of a shepherd named Gyges who got a magic ring that would make its wearer invisible and thus would enable him to go anywhere and do anything without being detected. Using this power of the ring, Glaucon entered the Royal Palace, killed the king, seduced the queen, and finally seized the throne. Now, Glaucon argues that if two persons – a rogue and a man of virtue – are each given such a ring, then the rogue would definitely use his ring to do all sorts of just and unjust acts as he pleased without being detected and punished. He also argues that the so-called virtuous man would also use the ring to best satisfy his self-interests, and so would do no better than the rogue. Moreover, none would have any moral qualms or strength of mind to keep him aloof from harming others for the sake of gratification of his desires.

Thus, the two views, viz., that each person by nature pursues his best interests, and that for each person the morally right action is the one that best promotes the agent's own interests suggested by Glaucon are known as *psychological egoism* and *ethical egoism*, respectively (Rachels, 2003). *Psychological egoism* may be defined as the *general factual claim* that human nature is so constituted that each person always chooses the best possible means of maximizing his own advantage or self-interest. *Ethical egoism* may be defined as the normative claim that the morally right or obligatory action is the one that best promotes the agent's own interests. Thus, while psychological egoism tells us how people do in fact behave, ethical egoism tells us how people ought to behave.

Now, if we assume that egoism and altruism are relative attributes that can vary on a continuous spectrum, then psychological egoism and ethical egoism may be

regarded as taking an extreme position. Some may find out obvious problems to these theories. But before going on to examine these two theories, let us just point out that if we take the other extreme position by defining *psychological altruism* as the factual claim that all humans by nature always seek to promote the welfare of others only and *ethical altruism* as the normative claim that all humans ought always to promote the welfare of others only, then these two theories would be in the same boat as the other two theories from the logical point of view and even more objectionable from the commonsense point of view.

Having clarified the relevant concepts, we will now briefly examine the Egoism-Altruism Debate with special reference to the issue as to whether *ethical egoism* can or cannot be sustained as a satisfactory universal theory of moral conduct. This depends on whether it passes the test of a good theory. Now, a theory is a good theory if and only if it satisfies two different general criteria, viz., consistency and completeness. The ***consistency condition*** requires that a good moral theory must be free from internal inconsistency and incompatibility with any reasonable moral intuition. Thus, for example, a system of folk morality containing two imperatives such as "Haste makes waste" and "The race is to the swift" involves an inconsistency and would, therefore, be unsatisfactory. The **completeness *condition*** requires that a good moral theory must be such that there should be no moral truth which is not derivable from the basic principles of the theory.

Let us first examine *psychological egoism* and see if it can be maintained as a universal theory. We will first consider the most common argument used to support

psychological egoism and then the objections against this theory (See, for example, Rachels, 2003).

Argument for Psychological Egoism:

Premise 1: Everyone does what they want to do.

Premise 2: What we want to do is always in our self-interest.

Conclusion: Therefore, we always do what is in our own self interest.

The above argument consists of two premises followed by the conclusion which is the claim made by psychological egoism. However, there are two serious objections against this argument as follows:

(1) Fallacy of Hasty Generalization (HG):

The Fallacy of Hasty Generalization usually occurs when the generalization is made on the basis of a selected sample which is not justifiably representative of the whole group. One objection against psychological egoism as an empirical theory is that it commits the fallacy of hasty generalization. The factual claim of this theory that all people act from the motive of self-interest is obviously false because there are many *disconfirming instances* against it. Thus many people act against their self-interests for various reasons such as, bad habits (e.g., smoking), religious beliefs, conscience, and so on.

(2) Fallacy of Unfalsifiability:

Falsifiability is one of the most important attributes of any scientific hypothesis. The Fallacy of Unfalsifiability occurs when one makes a general claim for which there is no possible way to prove it false. Another objection against psychological egoism is that it commits the Fallacy of Unfalsifiability by defining the theory in such a way that it defines any voluntary action as a self-interested action and thereby rules out all possible counterexamples to the theory and makes it a tautology that is empty of any factual content. Thus, for example, if someone donates blood to save the life of a serious patient who desperately needs blood, a thoroughgoing psychological egoist would interpret this to be a self-interested action motivated by the actor's desire for either getting pleasure or proving himself to be a hero or going to heaven. In fact, the very strategy of making the theory absolutely strong by turning down all possible counterexamples to it leads to its own ruin.

Ethical egoists sometimes claim that ethical egoism can be derived from psychological egoism. The reason for this is that if, according to psychological egoism, people always seek their own interests, then it would be useless to ask people to do otherwise. But obviously ethical egoism does not follow from psychological egoism. Moreover, if people do necessarily seek to promote their own interests, then it would be unnecessary and pointless to say that people ought to promote their own interests.

There is no serious argument for ethical egoism that deserves our careful attention probably because ethical egoists think that their position is unquestionably true.

Following Rachels (2003) we construct the most common argument for ethical egoism as follows:

Argument for Ethical Egoism:

(1) Each of us knows our own needs and wants better than anyone else, and so is better able to pursue our own wants and needs.

(2) We know the needs and desires of other people imperfectly and are not well situated to pursue them.

---

(3) Therefore, if we want to help others for which are not well suited, we would end up doing more harm than good to others.

---

(4) Therefore, the best way to help others would be for each to pursue our own interests.

By saying that we should not try to do things for which we are not competent so that we can avoid doing harm to others, this argument presupposes that we have a duty to help or not to harm others. Thus this argument for ethical egoism is self-defeating as it goes against ethical egoism.

From the above review of the *egoism-altruism dichotomy* discussed in chapter 2 is *unsustainable*, because neither universal egoism nor universal altruism is justifiable either as an empirical or a normative principle. It is interestingly to note that exactly the same two objections that have been raised against psychological egoism can also be

raised against *psychological altruism* as an empirical claim. People can behave egoistically at some times and altruistically at other times. In fact, even standard economic models of rational behavior do not require that men are essentially selfish in the sense in which it is interpreted in the egoism-altruism debate.

2.8 Conflict and Cooperation

Conflict and cooperation, as stated earlier, are two extremely important, pervasive and yet opposite forms of social interactions that bring about major changes in society. As we have already seen, game theory is the only discipline which is exclusively devoted to study of conflict and cooperation. We have also seen that the interests of any group of people involved in strategic interactions may be completely conflicting, wholly identical, or overlapping (i.e., mixed) which are studied by zero-sum games, pure coordination games, and mixed-motive games, respectively. It is, therefore, possible to define a continuum of conflicting situations varying from pure cooperative ones to pure conflict ones. We will now briefly discuss the nature and types of conflict in Subsection 2.8.1 and the nature and types of cooperation in Subsection 2.8.2.

2.8.1 Nature and Types of Conflict

The word “conflict” is used in different but related senses that are clearly reflected in the dictionary definitions. Thus conflict has been variously defined as “A prolonged armed struggle”, “A state of mind in which a person experiences a clash of

opposing feelings or needs", and "A serious incompatibility between two or more opinions, principles, or interests". A conflict may arise due to incompatibility of goals or incompatibility of means. Conflicts are not necessarily bad or god. It may have positive or negative impacts. A moderate amount of conflict often contributes toward a healthy organizational life. Conflicts may be just managed or resolved in a peaceful or a violent manner.

Conflicts may be classified into four different categories, interpersonal conflict, intrapersonal conflict, intergroup conflict, and intragroup conflict.

Interpersonal conflict refers to the conflict between two persons. The reason for such conflict is that everyone is a unique person and so people differ from one another in so many ways. Different people have different personalities which naturally lead to incompatible beliefs, desires, interests, and choices.

Intrapersonal conflict is a type of conflict that occurs within an individual due to conflicting internal psychological states, beliefs, thoughts, values, emotions, and principles and that may coexist in an individual. It may turn into frustration or severe depression that may require medical or psychiatric attention. There are different kinds of intrapersonal conflicts, such as approach-approach conflict, avoidance-avoidance conflict, approach-avoidance conflict, multiple approach-avoidance conflict depending on how a person is simultaneously pulled or pushed two or more different goals (Morgan, King, Weisz, & Schopler, 1986).

2.8.2 Nature and Types of Cooperation

It surprising that Tuomela (2000) is the first and as of now perhaps the only modern scholar who has tried to develop a comprehensive theory of cooperation. It is helpful to begin with the dictionary definitions of cooperation. The *Oxford Dictionary of English* defines cooperation as "The action or process of working together to the same end". The Collins English Dictionary defines "*Cooperation* is the action of working together with or helping someone". Obviously, the second definition mentions of a different type of cooperation that takes place when one person helps another person to achieve his goal and the two parties do not have any shared goal.

Tuomela (2000) defines cooperation as a joint action by a number of persons intending to achieve a shared goal. This definition implies that for an act of cooperation to take place, four conditions must be fulfilled. These are (i) a shared or common goal, (ii) joint action by the cooperators, (iii) the intentional action, and (iv) a number of persons. It may be noted here that Tuomela's definition does not cover a kind of cooperation involving no shared goal, no joint action, but intention and unilateral help from one party to second to help the latter achieve his goal. This type of cooperation has been recognized by the Collins English Dictionary. Tuomela divides cooperation into two types, viz., (i) *g-cooperation* or group mode cooperation and (ii) *i-cooperation* or individual mode cooperation. G-cooperation is based on a 'shared collective goal', but i-cooperation is based on 'compatible private goals'.

CHAPTER THREE

METHODOLOGY

3.1 Preview

Having explored and examined the literature on cooperation, we, in this chapter, look for and explain a suitable method of answering our research question. Since our inquiry as to why agents cooperate is a qualitative question which is mainly concerned with qualitative rather than quantitative data, the proper method of answering the question would be logical analysis and appraisal of the relevant concepts, statements, and arguments. Before embarking on a discussion of the main topics of the chapter, we point out the semantic distinction between a method of

inquiry and methodology which is relevant to the title of this chapter. What we use to answer our research question is a particular *method* and not *methodology*. A research method refers to the particular rules and techniques used for conducting a research. But research methodology is a wider concept and a discipline in its own right that deals in general with the various aspects and problems of research and particularly with the logic for the choice of an appropriate method from a number of available alternatives (Kothari, 2012). To defend our position against the most prevalent view about the reason for the occurrence or nonoccurrence of cooperation, we will now explain the method we are going to use and the reason for the choice of this method.

Below we discuss two different groups of concept – (a) different kinds of proof, as opposed to disproof, in logic, mathematics, and empirical sciences, and (b) distinction between conditions, causation and correlation, and sorting out several alternative conceptions of cause defined in terms of the distinction between necessary and sufficient condition.

3.2 Proof and Disproof: Direct vs. Indirect

The term ‘proof’ and its opposite ‘disproof’ are used not only in science and philosophy but also in everyday language. *The Merriam-Webster Dictionary* (2020) defines ***proof*** as ‘the process or an instance of establishing the validity of a statement especially by derivation from other statements in accordance with principles of reasoning.” We may formally define proof as the process or the product of the process of establishing the truth of a statement that logically follows from a set of given statements usually via an intermediate statement or sequence of statements where each statement follows from one or more preceding statements and the last statement in the whole sequence is the conclusion. In other words, proof consists in showing that the conclusion, i.e., the statement to be proved, follows from the premises, i.e., the supporting

statements, or that if the premises are true, the conclusion must be true. To sum up, proof is the process of establishing the truth of a statement.

Now, proof and disproof are concerned with the truth or falsity of statements. But a statement may be either *true*, such as (X+5)> (X+3), or *false*, such as $(a^2–b^2) \neq [(a+b)(a–b)]$, or a mere *conjecture* whose truth-value is unknown, i.e., undetermined, such as "Any even number greater than 2 is the sum of two primes" known as Goldbach's conjecture. It is important to note that that we cannot prove a false statement, because we can only disprove it. What we can do to show that a statement is false is to prove that its negation is true. Thus for any statement *P*, if we want to disprove *P*, we have to prove ¬*P*, (read 'not *P*'), because ¬*P* is true if and only if *P* is false. Thus ***disproof*** of any statement *P* may be defined as the proof of the negation of *P*.

There are various types of proof based on the nature and types of statement that we want to prove. But we will limit our discussion to a distinction between direct and indirect proof, because we will construct two indirect proofs to establish two different claims in chapter 5.

Proofs may be either *direct* or *indirect*. Put simply, a direct proof is one in which the conclusion is deduced from the given assumptions in a step by step fashion. Formally, a ***direct proof*** is one in which from an assumed statement *P* is deduced a second statement *Q* through the use the rules of inference, rules of replacement, axioms, and definitions. According to Keef and Guichard (2020), "A direct proof is a sequence of statements which are either givens or deductions from previous statements, and whose last statement is the conclusion to be proved" (p. 36). For example, let us consider the following *direct proof* of validity (i.e., formal truth) of an argument in symbolic logic:

1. $A \rightarrow C$	Premise	
2. $B \rightarrow C$	Premise/ $\therefore (A \vee B) \rightarrow C$	Conclusion
3. $\neg A \vee C$	1, Implication	
4. $\neg B \vee C$	2, Implication	
5. $(\neg A \vee C) \wedge (\neg B \vee C)$	3,4, Conjunction	
6. $(\neg A \wedge \neg B) \vee C$	5, Distribution Law	
7. $\neg (A \vee B) \vee C$	6, De Morgan's Law	
8. $(A \vee B) \rightarrow C$	7, Implication	

The above proof for the argument consists of eight statements numbered sequentially from (1) to (8) where statements (1) and (2) are the premises from which is derived statement (8) which is the last one in the sequence and the conclusion, while the justification for each statement, including the intermediate ones, in the proof is written on the right side of it.

Unlike a direct proof, an ***indirect proof*** is concerned not with deriving the conclusion from the assumptions to show that it is true, but rather with proving it indirectly by showing that its denial leads to a contradiction so that it cannot be false. An indirect proof is a subtle and ingenious technique of proof that can always be used including the cases where direct proof is inapplicable. Hence, an indirect proof is more powerful than a direct proof. That the use of indirect proof is common even in our ordinary life is evident from an example from Keef and Guichard (2020). Suppose, you want to know whether the sky is sunny or overcast now, because you are sitting in a room and cannot see through the window. However, you can *indirectly* know about that from the quality of light you can see from your room.

There are two kinds of closely related indirect proof, viz., proof by contraposition and proof by contradiction. They are indirect proof in the sense that in both of them the proof procedure starts by assuming the negation of the conclusion.

The ***proof by contraposition*** is based on the principle, [(P⇒Q)⇔(¬Q⇒¬P)], that is, a conditional statement (P⇒Q) is logically equivalent to its contrapositive (¬Q⇒¬P). Since these two conditionals are logically equivalent, proving any one of them would tantamount to proving the other. Proving (¬Q ⇒¬P) is often easier than proving (P⇒Q). In a proof by contrapositive, we first assume that Q is false and from this we infer that P is false. Thus, for example, it is more natural and easier to prove the statement "If this is not red, then this is not my pen" than its equivalent "If this is my pen, then this is red".

Hammack (2018) informally but briefly described the idea of a ***proof by contradiction***, also known as *proof by reductio ad absurdum*, as follows:

> The basic idea is to assume that the statement we want to prove is false, and then show that this assumption leads to nonsense. We are then led to conclude that we were wrong to assume the statement was false, so the statement must be true (p. 137).

A *proof by contradiction* of a theorem, such as $P \Rightarrow Q$, is one where the conclusion Q is not derived from the assumption P but rather the negation of *Q*, i.e., ¬*Q* is assumed as true and added as a conjunct to *P* to form ($P \wedge \neg Q$) as the new premise from which is derived an explicit contradiction (r∧¬r) that gives a reason to hold that *Q* cannot be false, and, therefore, *Q* has been proved to be true. The reason for the truth of *Q* is not difficult to see. If P and ¬*Q* are both true, then ($P \wedge \neg Q$) must also be true. If ($P \wedge \neg Q$) is true, then it cannot logically imply any false

statement, such as (r∧¬r). But the fact that the statement ($P \wedge \neg Q$) logically implies the necessarily false statement (r∧¬r) means that ($P \wedge \neg Q$) must be false and not true. Since the conjunct P is a given truth, the only possibility to make ($P \wedge \neg Q$) false is to accept that Q is true. Formally, $[(P \wedge \neg Q) \Rightarrow (r \wedge \neg r)] \Rightarrow Q$. This proves the theorem $P \Rightarrow Q$.

To explain the concept of *indirect proof* (Hurley, 2012) let us now consider the following example:

1. $p \rightarrow q$ Premise
2. $p \vee q$ Premise/ ∴ q Conclusion
3. ¬q AIP (Assumption for Indirect Proof)
4. ¬p 1,3, Modus Tollens
5. q 2,4, Disjunctive Syllogism
6. $q \wedge \neg q$ 5,3, Conjunction
7. q 3-6, IP (Indirect Proof)

The above proof for the argument consists of seven statements numbered sequentially from (1) to (7) where the justification for each statement in the proof is written on the right side of it. Here statements (1) and (2) are the given premises. Statement (3) is the one that we want to show follows from the premises and hence has been assumed as false to serve as an *assumption for indirect proof* (AIP). The indirect proof sequence begins with the AIP at line (3) and ends at line (6) where the AIP leads to an explicit contradiction that entitles us assert the denial of the AIP as the conclusion in line (7).

It may be pointed out that the method of indirect proof has at least two important advantages over direct proof. First, there is *no direct proof* for some arguments, such as those like

$$A / \therefore [B \vee (B \rightarrow C)]$$

whose conclusion is a tautology (Copi, 1996, pp. 53-54). Second, the indirect method of proof for validity of argument is often more efficient than the direct method as the former can be completed more quickly and in a fewer number of steps than the latter (Copi et al, 2014, p. 421).

3.3 Causation, Conditions, and Correlation

Knowledge of causal connection between *cause* and *effect* as two different events or conditions where the latter invariably or at least sufficiently frequently follows the former enables us to explain and predict natural phenomena and discover laws of nature which in turn allow us to gain enormous control over our environment. The concept of causality or causation may be approached from two different perspectives. The relation between cause and effect may be considered as either a sort of *relation* between a necessary condition and a sufficient condition which goes from one to the other, or a *correlation* that measures the strength of positive or negative linear relationship between two variables.

Let us now consider the concept of causation in terms of necessary and sufficient conditions. The distinction between these two types of conditions may be explained through a conditional which is a compound statement of the form "If P, then Q" denoted symbolically by $(P \rightarrow Q)$ where P is called the antecedent (or, hypothesis) and Q is called the consequent (or, conclusion). It is to be noted here that $(P \rightarrow Q)$ is equivalently read as 'P *implies* Q', '*If* P *then* Q', 'P *only if* Q', 'P is a *sufficient* condition for Q', and 'Q is a *necessary* condition for P'.

Before defining the two related concepts of necessary condition and sufficient condition, let us look at the distinction between them with the help of a conditional statement, such as '*If* Mary is a mother, *then* Mary is a woman'. Here the antecedent 'Mary is a *mother*' is a sufficient condition for the consequent 'Mary is a *woman*', because 'being a mother' is sufficient for 'being a woman'. On the other hand, the consequent 'Mary is a *woman*' is a necessary condition for the antecedent 'Mary is a *mother*', because 'being a woman' is absolutely necessary for 'being a mother', i.e., one *cannot* be a mother without being a woman first. It is worth noting that 'being a mother', though sufficient, is *not necessary* for 'being a woman'. Similarly, 'being a woman', though necessary, is *not sufficient* for 'being a mother'.

Having looked at an example, we will now define the concepts of necessary condition and sufficient condition. It must, however, be noted that these two concepts may be alternatively defined in terms of either events, or statements, or even state of affairs, but the definitions remain essentially the same (See Swartz, 1997; Copi and Cohen, 2004). We define the concepts below first in terms of events to emphasize the relevance of the concept of causation for our purpose, and then in terms of statements or propositions for the sake of analytical convenience. In stating the definitions we will use the following shorthand symbols in addition to the customary symbols of logic:

NC	denotes necessary condition
SC	denotes sufficient condition
NSC	denotes necessary and a sufficient condition
$=_{df}$	denotes equal by definition
$PC^{N}Q$	denotes P is a necessary condition for Q
$PC^{S}Q$	denotes P is a sufficient condition for Q

$PC^{N\wedge S}Q$ denotes P is a necessary and a sufficient condition for Q

$PC^{N\wedge\sim S}Q$ denotes P is a necessary but not a sufficient condition for Q

$PC^{\sim N\wedge S}Q$ denotes P is not a necessary but a sufficient condition for Q

$PC^{\sim N\wedge\sim S}Q$ denotes P is neither a necessary nor a sufficient condition for Q

In terms of Events:

E_1 is a NC for $E_2 =_{df}$ If event E_1 is *absent* i.e., does not happen, E_2 cannot happen.

E_1 is a SC for $E_2 =_{df}$ If event E_1 is *present* i.e., happens, E_2 must happen.

In terms of Statements:

A statement *P* is a NC for another statement *Q* if and only if the falsity (or, denial) of *P implies* (or, guarantees) the falsity of *Q*. Symbolically, $\mathbf{PC^{N}Q =_{df} (\neg P \rightarrow \neg Q)}$.

A statement *P* is a SC for a statement *Q* if and only if the truth of *P implies* (or, guarantees) the truth of *Q*. Symbolically, $\mathbf{PC^{S}Q =_{df} (P \rightarrow Q)}$.

Thus, for example, being female is a *necessary* condition for being a mother, because one cannot be a mother without being female, i.e., (¬Female → ¬Mother). On the contrary, being a mother is a *sufficient* condition for being female, because if it is true that a person is a mother then it must also be true that that person is a female. That is, (Mother → Female). Similarly, consider the conditional statement 'If there is *life*, there is *oxygen*' that expresses a causal connection. It shows that the antecedent 'There is life' is

a SC for the consequent 'There is oxygen' just because the *presence* of life is sufficient to guarantee the existence of oxygen, while the consequent is a necessary condition for the antecedent just because the *absence* of oxygen is necessary to guarantee the nonexistence of life.

It is noteworthy that NC may be either ***necessarily*** NC, i.e., necessary from the logical point of view of *logic* or ***contingently*** NC, i.e., necessary as a *matter of fact*, and so, SC, being the converse of NC, may also be either ***necessarily*** SC or ***contingently*** SC (Wolfram, 1994). For example, in the conditional "If Nancy has *brown eyes*, then Nancy has *eyes*" having 'brown eyes' is a **necessarily SC** for having 'eyes', and so having 'eyes' is a **necessarily NC** for having 'brown eyes'. But in the conditional "If A *drinks hemlock*, then A *dies*" there is a *causal* and *contingent* connection between the antecedent and the consequent, and hence 'drinking hemlock' is a causally and ***contingently*** SC for 'death' and 'death' is a causally and ***contingently*** NC for 'drinking hemlock'.

Now, there are *five* important *features* of the relation between NC and SC. First, necessary condition is defined in terms of ***absence***, but sufficient condition is defined in terms of ***presence***. Second, NC and SC are ***converses*** of each other. Thus, in $(P \rightarrow Q)$, P is a SC for Q and Q is a NC for P. On the other hand, in $(Q \rightarrow P)$, P is a NC for Q and Q is a SC for P. Third, any conditional statement, such as $(P \rightarrow Q)$ is logically equivalent to its ***contrapositive*** $(\neg Q \rightarrow \neg P)$ and not to its converse $(Q \rightarrow P)$; but $(Q \rightarrow P)$ is logically equivalent to its own contrapositive $(\neg Q \rightarrow \neg P)$. Therefore, the definition of "P is a necessary condition for Q" may be written either as $(\neg P \rightarrow \neg Q)$ or as $(Q \rightarrow P)$.

Thus the universal statement "All squares are quadrilaterals" which can be symbolized as '$(\forall x)(Sx \rightarrow Qx)$' is equivalent to its contrapositive "All non-quadrilaterals are non-

squares" symbolized as '$(\forall x)(\neg Qx \rightarrow \neg Sx)$', and not to its converse "All quadrilaterals are squares" symbolized as '$(\forall x)(Qx \rightarrow Sx)$'. Fourth, the logic of necessary condition corresponds to ***Modus Tollens***, while the logic of sufficient condition corresponds to ***Modus Ponens***. Thus for example, in $(P \rightarrow Q)$ if Q is a necessary condition for P, and Q is false, then P is false is fundamentally similar to $[\{(P \rightarrow Q) \wedge \neg Q\} \rightarrow \neg P]$. On the contrary, if P is a sufficient condition for Q and if P is true, then Q is true is similar to Modus Ponens: $[\{(P \rightarrow Q) \wedge P\} \rightarrow Q]$. Fifth, there may be alternative sets of SCs that may yield the same NC, and different sets of conditions may be NC for the same set of SC. These are true of both mathematical statements as well as empirical statements of causal connection. Thus, the same NC such as (x>z) logically follows from three different SCs such as $[(x>y) \wedge (y>z)]$, $[(x=y) \wedge (y>z)]$, and $[(x>y) \wedge (y=z)]$. On the other hand, the same SC such as $[(w>x) \wedge (x=y) \wedge (y>z)]$ may have more than one NCs such as (w>y) and (w>z). In causal conditionals there may be many alternative SCs, such as 'drinking hemlock' and 'poisonous snake bite' that may be the cause of 'death' as a NC. On the contrary, giving the same drug to two patients with the same disease may produce recovery from sickness for one but serious side effect for the other.

However, true conditional statements of the form "If P, then Q" symbolized as "$(P \rightarrow Q)$" may express various types of ***implication*** depending on the meaning and nature of the relation that holds between its antecedent *P* and consequent *Q* (Copi & Cohen, 1997). The meaning of each of the four different types of implication listed below can be easily grasped from the following *examples*:

(1) Logical Implication: If all judges are wise and Tom is a judge, then Tom is wise.

(2) Definitional Implication: If Dick is a father, then Dick is a male person.

(3) Causal Implication: If you strike a match, then it will light.

(4) Decisional Implication: If you pass the examination, I will present you a gift.

A quick glance would show that each of the above conditional statements from (1) to (4) asserts a distinct type of real implication relation between the antecedent and the consequent which makes the particular conditional true. These conditionals express the relationship of *logical implication*, *definitional implication*, *causal implication*, and *decisional implication* respectively. But each of these conditional statements carries an additional meaning that happens to be the partial common meaning of each to them. Thus each conditional of the form "If P, then Q" asserts *a truth-functional relation* between its antecedent and its consequent to the effect that "If the antecedent P is true, then the consequent Q must be true as well". In other words, "If P, then Q" asserts that "It is not the case that P is true but Q is false" which may be symbolized as $\neg(P \wedge \neg Q)$.

Now, let us isolate '$\neg(P \wedge \neg Q)$' which is the truth-functional part of the meanings of the various conditional statements of the form "If P, then Q" and consider it as the sole meaning of a distinct but weak type of implication relation that may be symbolized by the arrow symbol "$\rightarrow$". Therefore, this particular sort of conditional statement expressing the truth-functional implication relation between any statement P and any statement Q may be symbolized as '$(P \rightarrow Q)$'. This truth-functionally conditional statement '$(P \rightarrow Q)$' is known as 'material implication', a term coined by Russell (1903), and may be listed as the fifth type of implication as follows:

(5) Material Implication: If London is the capital of Germany, then 2+3=10.

The partial common meaning of the various types of conditional statements captured by '¬(P∧¬Q)' which is a truth-functionally compound statement and abbreviated as 'P → Q' may be defined by the following truth table (Copi & Cohen, 2004):

TABLE 3.1 PARTIAL MEANING OF *CONDITIONAL STATEMENTS*

1	2	3	4	5	6
P	Q	¬Q	P∧¬Q	¬(P∧¬Q)	P → Q
T	T	F	F	T	T
T	F	T	T	F	F
F	T	F	F	T	T
F	F	T	F	T	T

Now, in Table 3.1 'P → Q' is equivalent to '¬(P∧¬Q)' by definition. So, if we now remove '¬(P∧¬Q)' from the table, we can delete column 5, and consequently we can also remove 'P∧¬Q' and '¬(P∧¬Q)' and delete columns 3 and 4 that were required as an aid to constructing the table. The full meaning of material implication as a separate but fundamental type of implication that conveys only the partial common meaning of all the four other types of implication may now be defined by the following truth table (Copi & Cohen, 2004):

TABLE 3.2 DEFINITION OF *MATERIAL IMPLICATION*		
P	Q	$P \rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

Let us now consider statement (5) above which is a conditional statement where the antecedent P is "London is the capital of Germany" that happens to be false and the consequent Q is "2+3=10" that is known to be false. But the fourth and last row of Table 3.2 shows that '$P \rightarrow Q$' must be true when P and Q are both false. Hence, the conditional statement 'If London is the capital of Germany, then 2+3=10' must be accepted as true, although there is no real connection between 'London is the capital of Germany' and '2+3=10'. One reason for accepting this strange eventuality is that we reasonably agreed to accept '$\neg(P \wedge \neg Q)$' as a definition of '$P \rightarrow Q$'. Another reason is that if we refuse to accept this definition and try to adopt any alternative but truth-functional definition of '$P \rightarrow Q$', we will be forced to accept even more surprising and stranger consequences (Anwar, 1996).

Now, it is extremely important to understand that in any given conditional ($P \rightarrow Q$), P and Q may be related to each other in ***four*** possible ways each of which represents a distinct but relevant way of interpreting the concept of *cause*. Let us first try to explain them with examples:

P is necessary but not sufficient for Q: Being a parent is necessary but not sufficient for being a father. The reason for this is that to be a father a person must not only be a parent but also be a male and a mother is also a parent.

P is sufficient but not necessary for Q: In the statement "If something is red, then it is colored" being red is sufficient but not necessary for being colored. The reason is simply that whatever is red is colored, but whatever is colored need not necessarily be red.

P is both necessary and sufficient for Q: Being frozen water is necessary and sufficient for being ice. The reason for this is that if something is frozen water it must ice and moreover if it is not frozen water, i.e., either water but not frozen or something frozen but not water, then it cannot be ice.

P is neither necessary nor sufficient for Q: Being 6 feet tall is neither necessary nor sufficient for winning a lottery. Obviously, being 6 feet tall is not necessary for winning a lottery, for one may win it without being exactly 6 feet tall. Again, being 6 feet tall is not a sufficient condition, for it does not guarantee winning a lottery prize that happens to be a purely random event.

Let us now take a look at the views of some prominent philosophers on the nature of the relationship between cause and effect. Now, there seem to be no doubts among philosophers that cause and effect as events or conditions do not simply happen, they happen in *time*, one after

another. Moreover, they occur not at random, but in a definite order where the cause *precedes* the effect and never in reverse order. But there is a sharp disagreement among scholars as to whether the *relation* whereby the cause precedes the effect, or the converse relation through which the effect succeeds the cause, is invariable (i.e., necessary) or variable (i.e., contingent). For example, Mill (1882) holds that the cause and the effect are "invariably and unconditionally" connected, while al-Ghazali (2000) and Hume (1740/1888) offer compelling arguments for the opposite view that that the relation between cause and effect is contingent and not necessary.

Mill (1882) held that there is a necessary connection between the *cause* as an antecedent or set of antecedents and the *effect* as a consequent such that the effect "invariably and unconditionally" follows from the cause, and so we can infer the cause from the effect and the effect from the cause by the use of experiments and the rules of deduction. As Mill (1882) writes:

> We may define, therefore, the cause of a phenomenon, to be the antecedent, or the concurrence of antecedents, on which it is invariably and unconditionally consequent. Or if we adopt the convenient modification of the meaning of the word cause, which confines it to the assemblage of positive conditions without the negative, then instead of "unconditionally," we must say, "subject to no other than negative conditions." (p. 418).

For Mill, positive conditions are those antecedents that lead to the production of the consequent or effect, while negative conditions refer to the absence of those circumstances that prevent the effect from taking place.

But Al-Ghazali (2000) convincingly argues against the belief of the so-called necessary connection between cause and effect thus:

> The connection between what is habitually believed to be a cause and what is habitually believed to be an effect is not necessary, according to us. But [with] any two things, where "this" is not "that" and "that" is not "this" and where neither the affirmation of the one entails the affirmation of the other nor the negation of the one entails negation of the other, it is not a necessity of the existence of the one that the other should exist, and it is not a necessity of the nonexistence of the one that the other should not exist—for example, the quenching of thirst and drinking, satiety and eating, burning and contact with fire, light and the appearance of the sun, death and decapitation […] and so on to [include] all [that is] observable among connected things in medicine, astronomy, arts, and crafts. Their connection is due to the prior decree of God, who creates them side by side, not to its being necessary in itself, incapable of separation (p. 166).

Thus Al-Ghazali virtually subscribes to the view that the so-called cause is neither a necessary nor a sufficient condition for the production of the effect.

We will now take a look at how the cause-effect relation known as *causation* or causality differs from *correlation*. Causation may be defined as the impact or power by which one event or set of conditions called cause contributes to the occurrence of another event or set of conditions called effect. For example, a certain kind of bacteria or virus may, under certain conditions, act as a cause of a particular type of disease. Correlation, on the other hand, refers to a measurement of the strength or degree of positive or negative relationship between a pair of variables that are linearly related and change together. It just says how two things vary together, but by itself does not tell us why they vary together at all. For example, there is a *positive correlation* between the *sale of ice creams* and the *sale of sunglasses*, because the sales of the two things either increase or decrease together. To take another example, there is a *negative correlation* between the *sale of ice creams* and the *sale of warm clothes*, because the sales of the two things are inversely related which means that if there is an increase in the sale of one item, there is a corresponding decrease in the sale of the other item, and vice versa. Since correlation is only a measure of the linear positive or negative relationship or non-relationship between two variables regardless of any consideration

of which variable is dependent and which one is independent, it cannot identify one variable as the cause and the other as the effect. If we look beyond correlation, we can see that in both the cases the underlying real causal factor is the change in *weather temperature* that can explain the magnitude and direction of relation between the *sale of ice creams* and the *sale of sunglasses* on the one hand and those between the *sale of ice creams* and the *sale of warm clothes* on the other.

Now, in view of the fact that causation may be conceived of as a kind of connection between cause and effect as two different conditions or states of affairs where the latter constantly follows the former, we can specify certain important features of it. First, causality is essentially a connection or *relationship*. Second, it is a *dyadic* relation between two different occurrences or conditions. Third, causality implies a *temporal order* in so far as the cause happens before and not after or simultaneously with the effect. Fourth, the order is constant or sufficiently frequent that makes the causal relation lawful or at least lawlike. Fifth, causality is an *irreflexive* relation, because nothing can be its own cause. Sixth, causality is an *asymmetric* relation, because, given any two events *C* and *E*, if *C* is a cause of *E*, then *E* is not a cause of *C*. Seventhly, if there is a *causal sequence* of several events where A causes B, B causes C, C causes D, and D causes E, E may be regarded as the effect of any one or all of the events. Of all the preceding events D is the nearest one and hence regarded as the *proximate cause* of E, while other ones from C to A are more and more *remote causes* of E.

However, it cannot be denied that the word "cause" is used in different senses in different contexts (Hatcher et al., 1990; Copi & Cohen, 2004; Copi, Cohen, & McMahon, 2014). First, it is sometimes used in the sense of NC when the intention is to prevent or *eliminate* some undesirable phenomena, such as a disease, by discovering some essential factor, such as the germ that causes the disease. When we want to infer cause from effect we use cause only in the sense of NC. Second, sometimes the word

"cause" is used in the sense of SC when the intention is to *produce* something desirable rather than to *eliminate* something undesirable. Thus a doctor may prescribe a number of measures, such as diet, medicine, and exercise in order to produce good health. When we want to infer effect from cause we use cause only in the sense of SC. Third, the word "cause" is used in the sense of necessary and sufficient condition when inferences are drawn from cause to effect as well as from effect to cause. This usage of cause is identified with the SC which is regarded as the conjunction of all NCs. This conception of cause implies that there is a unique cause of every event and hence it is opposed to the doctrine of plurality of causes.

An elusive and yet important sense of the word cause recognized especially by medical scientists is that cause is neither necessary nor sufficient but still a 'contributing condition' or 'contributory cause' of many events or effects. For example, smoking is definitely not a sufficient condition of lung cancer, because many people continue to smoke without ever having cancer. But smoking is not a necessary condition of lung cancer either, because many people who never smoke have got such cancer. Nevertheless, smoking in combination with other relevant factors often plays a positive role in the production of lung cancer. It is, therefore, reasonable to infer that smoking, despite being neither a necessary nor a sufficient condition, is a relatively "weaker" but nonetheless "clinical cause" of cancer (Riegelman, 1979; Hatcher et al., 1990; Copi, Cohen, & McMahon, 2014; Kelley, 2014).

Using the conceptual tools discussed in this chapter, we will consider in chapter 5 the possible connection between altruism and cooperation and show that altruism is neither a necessary nor a sufficient but rather a 'contributing condition' for cooperation.

CHAPTER FOUR

THE *PRISONER'S DILEMMA* AND THE *CHICKEN*: A COMPARISON

4.1 Preview

In this chapter we discuss two extremely important and pervasive games known as the *Prisoner's Dilemma* and the *Chicken* each of which may be conceived of as a one-stage, two-person, two-strategy, simultaneous-move, non-zero-sum, non-cooperative, mixed-motive, and ordinally symmetric game. In each of these games each player is goal-

oriented and independently chooses a strategy such that the two players' strategy-choices jointly determine a unique Pareto-inefficient Nash equilibrium outcome in case of the former but two Pareto-efficient pure strategy Nash equilibria, one mixed strategy Pareto-inefficient Nash equilibrium, and a *correlated equilibrium* that is both a Pareto-efficient and neutral outcome in case of the latter. We also discuss the similarities and the differences between the games with respect to what the social value orientations of the players are, whether there is a unique or multiple Nash equilibria, whether the players can attain a dominant strategy equilibrium, whether there exists a correlated equilibrium, whether the Nash equilibrium attained is also a dominant strategy equilibrium, whether the Nash equilibrium attained is Pareto-inefficient, and whether it is possible to obtain a mutually more advantageous or better outcome.

4.2 The *Prisoner's Dilemma*

We will now present the well-known *Prisoner's Dilemma* (PD) game which is at the core of game theory and serves as a paradigm case by reference to which several strategies for dealing with numerous real life situations of conflict and cooperation can be explained. The PD may be described as an instance of a two-person, two-strategy, simultaneous-move, non-zero-sum, non-cooperative, and ordinally symmetric game. This means that the game has the following features: (1) The game is played between *two players* or parties. (2) Each side can independently choose any one from *two mutually exclusive strategies* or courses of action, viz., cooperation and defection (i.e., non-cooperation) (3) The two sides make their own choice of action *simultaneously* and not sequentially, i.e.,

at the same time and not one after another, so that each party's decision is made merely on the basis of a guess and not knowledge about what the other party's choice is going to be. (4) The game is *non-zero-sum* in the sense that the payoffs of the two players do not sum to zero so that the players' interests are not wholly opposed This means that one player's gain is not always at the loss of the other player, and hence some opportunity for mutual cooperation exists. (5) It is a *non-cooperative* game in the sense that the players make independent and strategic decision where there can be no "binding agreement" between them to enforce any action on one another. (6) The PD is a *mixed-motive* game in the sense that the interests of the players as reflected in the payoff numbers are neither diametrically opposed nor completely identical but rather mixed. This is why the players in such games are motivated partly to cooperate and partly to defect. (7) The game is essentially *ordinal* in the sense that players' utilities are assumed to be ordinal, i.e., they can only be ordered as more or less, and not cardinal, i.e., they are not amenable to mathematical calculation such as addition and subtraction. This means that the game is defined in terms of how the outcomes are ranked or ordered by the players in terms of their relative desirability. (8) The game is *symmetric*, because the players have the same available actions and when the players' positions are reversed the ordering of the outcomes are also reversed.

Now, at the heart of the *Prisoner's Dilemma* game lies a tension between collective rationality that recommends mutual cooperation and individual rationality that leads to mutual defection because of each party's temptation to defect against the opponent's cooperation, on the one hand, and the fear of turning out to be a sucker, on the other. The two-person PD may be stated as a situation involving two players (or, parties)

making interdependent decisions such that each person chooses between two strategies, viz., cooperate (C) and defect (D), each obtains a greater payoff from choosing D rather than C regardless of whether the other person chooses C or D, but each gets a better payoff if both choose C than if both choose D. Thus, though the Prisoner's Dilemma game is a common phenomenon in everyday life where two players have the option to cooperate for mutual benefit, yet each also faces the risk of being cheated.[14]

Although the *Prisoner's Dilemma* is usually attributed to Melvin Dresher and Merrill Flood who were working at RAND (Straffin, 1993), it was conceived of much earlier by others especially Thomas Hobbes in his *Leviathan* (1651), David Hume in his *A Treatise of Human Nature*, (1740/1888), and Giacomo Puccini, as pointed out by Hargreaves Heap and Varoufakis (2004, p.174.), in his famous opera of romance and tragedy, *Tosca* (1899). There is a clear expression of the essentials of the *Prisoner's Dilemma* game between two selfish persons arriving at a mutually disadvantageous outcome in Hume's (1740/1888) writing as follows:

> "Your corn is ripe to-day; mine will be so tomorrow. It is profitable for us both, that I should labour with you to-day, and that you should aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon

[14] It is important to note that when this argument is more formally stated as

Premise 1: D|C > C|C and D|D > C|D

Premise 2: C|C > D|D

Conclusion: D|C > C|C > D|D > C|D

where the symbolic expression 'D|C', for example, stands for the statement that 'one player defects while the other cooperates' and so on for the others, it may be intuitively understood to be logically valid even without a complete proof of formal validity that would require suitable interpretation of the symbols.

> your account; and should I labour with you upon my own account, in expectation of a return, I know I should be disappointed, and that I should in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security (*Treatise of Human Nature*, Bk. III, Part II, Sec. V.)."

But it was Alfred W. Tucker (1950), a Canadian born Stanford professor of mathematics and supervisor of John Nash's PhD thesis, who invented the story of the *Prisoner's Dilemma*, wrote a one page short note on it, named the game accordingly while presenting it in a seminar of the Psychology Department of Stanford University in 1950, and thereby popularized not only this particular game but rather game theory as a whole (Straffin, 1980, 1993).

A variant of the ***story*** of the *Prisoner's Dilemma* goes somewhat like this. The police arrests two persons, say, Alfred and Brown, on charge of a minor crime of carrying unauthorized guns. The smart district attorney, however, believes that they jointly committed a more serious crime of bank robbery, but has no sufficient evidence to prove their guilt. He can, however, convict them only if one or both confess their offence. So, he puts them under separate prison cells so that they cannot communicate with each other before interrogation. He then summons the prisoners individually and offers each of them the following deal:

(1) If you confess (i.e., disclose) and thereby implicate your partner, but he remains silent (i.e., conceals), then for testifying as state's witness against him you will go *free* (i.e., get a sentence of *zero years* in jail) while he will get the *maximal*

sentence of *five years*. On the contrary, if your partner confesses but you do not, then the jail terms will be reversed for you.

(2) If you both disclose and thereby implicate each other, both of you will get a *moderate* sentence of *three years* for such a serious crime in view of your confession.

(3) However, if you both conceal, both of you will get the *minimal* sentence of *one year* on the basis of the evidence for committing a much lighter crime rather than the more serious one that could not be proved.

It must be borne in mind that in this game to conceal is to *co-operate* since by jointly concealing the two parties can attain the best possible collective outcome which is one year for each prisoner. But to disclose is to implicate the other, and so is to *defect*, since to disclose is to betray the other prisoner from getting a lighter jail term.

The above-mentioned information may now be represented by the payoff matrix in Table 4.1 as follows:

TABLE 4.1: PAYOFF MATRIX FOR *PRISONER'S DILEMMA*

		Brown	
Alfred	Strategy	Cooperate (Conceal): C	Defect (Disclose): D
	Cooperate (Conceal): C	(-1, -1)	(-5, 0)
	Defect (Disclose): D	(0, -5)	(-3, -3)

It is important to note two things in Table 4.1. First, the first entry in each ordered pair of numbers in a cell denotes the row player's, i.e., Alfred's, payoff and the second entry denotes column player's, i.e., Brown's payoff. Second, the jail terms indicate disutility and hence are denoted by negative numbers. But in cases where the payoffs indicate utility gain they would be denoted by positive numbers. Now, if we substitute the numbers 5, 3, 1, and 0 for the four payoff numbers 0, -1, -3, and -5 respectively and thereby keep the relative ranking of the outcomes unchanged and take the Row player Alfred as Player I and the Column player Brown as Player II, then the ranking of the four possible outcomes by each player would remain unchanged and so the *Prisoner's Dilemma* could be represented by Table 4.2 as shown below.[15] The same result could also be obtained by first making a diagonal tradeoff of the outcomes and then multiplying each payoff number by -1.

[15] The PD can be more fully and formally stated as a *pure-strategy* analysis of *one-stage*, *two-person*, *two-strategy*, *simultaneous-move*, *non-zero-sum*, *non-cooperative*, and *ordinally symmetric* **game** in which each of the players has a *dominant strategy* which when chosen leads to a *unique Nash equilibrium* outcome that happens to be *Pareto-inefficient*.

TABLE 4.2: *PRISONER'S DILEMMA*: DEFINITION & PAYOFF MATRIX

(A Mixed-motive Game)

	Player II		
Player I	Strategy	Cooperate (C)	Defect (D)
	Cooperate (C)	(R, R) (3, 3)	(S, T) (0, 5)
	Defect (D)	(T, S) (5, 0)	(P, P)* (1, 1)*

Following Rapoport and Chammah (1965) and Axelrod and Hamilton (1981), the *Prisoner's Dilemma* may be defined as a game involving two players, player I and player II, each having two possible mutually exclusive choices, cooperate (C) and defect (D), so that the choices of the two players jointly determine the four possible outcomes, (C,C), (CD), (D,C), and (D,D) specified respectively as (R,R)=(3,3), (S,T) =(0,5), (T,S) =(5,0), and (P,P) =(1,1) in Table 4.2 by the four payoff numbers (*T*, *R*, *P*, *S*) = (5, 3, 1, 0) that satisfy the two conditions denoted by T>R>P>S and R>(T+S)/2 where T, R, P, and S for each player are defined as follows:

T = Payoff for (D|C), i.e., Payoff for Defection against Cooperation, or Temptation to Defect against Cooperation

R = Payoff for (C|C), i.e., Payoff for Cooperation against Cooperation, or Reward for Mutual Cooperation

P = Payoff for (D|D), i.e., Payoff for Defection against Defection, or Punishment for Mutual Defection

S = Payoff for (C|D), i.e., Payoff for Cooperation against Defection, or Sucker's Payoff for Cooperation against Defection

Table 4.2 as well as Table 4.1 represents a *game* in the sense that it is a situation that involves the following elements:

(1) There are at least two goal-oriented but interdependent decision makers called *players* – here Row Player known as Player I and Column Player known as Player II.

(2) Each player has to make a *choice* between two or more mutually exclusive but collectively exhaustive *alternatives* – here 'Cooperate' and 'Defect'.

(3) Each player's choice will be *optimal* in the sense that each will make the best possible choice of an action by pursuing his or her goal consistently.

(4) Each player's choice of an action is *strategic* in the sense that it is conditional upon the expected choice of action by the other.

(5) Each player receives a numerical *payoff* when the game is played by the players.

(6) Each possible ordered pair of payoff numbers, one for each player, is called an *outcome* determined by a given combination of strategies, one for each player. By

convention the first number in each ordered pair denotes Player I's payoff and the second denotes Player II's payoff.

Thus Table 4.2, for example, shows that for the two players each having two strategies, there will be four possible combinations of strategy – (Player I Cooperates, Player II Cooperates), (Player I Cooperates, Player II Defects), (Player I Defects, Player II Cooperates), and (Player I Defects, Player II Defects) – specifying the four corresponding outcomes – (R, R), (S, T), (T, S), and (P, P), or (2, 2), (0, 3), (3, 0), and (1, 1) – respectively each of which is an ordered pair of payoff numbers.

It is notable that the condition T>R>P>S is known as the *Ordering Condition* and the condition 2R>(T+S), or equivalently R>(T+S)/2, as the *Anti-exploitation Condition*. In *non-iterated* version of the *Prisoner's Dilemma* where the game is played only one time the ordering condition alone is used to define the game (Luce and Raiffa, 1957). The ordering condition denoted by T>R>P>S states that in a PD each player ranks his four payoffs in decreasing order of preference as T, R, P, and S. That means for each agent defection against cooperation yields a better payoff than mutual cooperation, mutual cooperation produces a greater payoff than mutual defection, and mutual defection gives a better payoff than cooperation against defection. But the anti-exploitation condition denoted by 2R>(T+S) is added to the ordering condition to define the *iterated* version which can be used to take care of evolutionary cases. The second condition states that the reward for mutual cooperation for two successive plays of the game has to be greater than the sum of the payoff for temptation and the sucker's payoff in order to make it possible

for mutual cooperation to be better than turn-taking in exploiting each other, i.e., exploiting and being exploited.

Now, if each player is assumed to be rational in the sense of seeking to maximize its own advantage by always preferring a higher payoff to a lower one, then the most preferred or best outcome for Row is (5, 0) which turns out to be the most dispreferred or worst outcome for Column. Thus, in descending order of preference Player I and Player II in Table 4.2 would rank the four outcomes differently as shown by Table 4.3 below:

TABLE 4.3: RANKING OF THE OUTCOMES BY PLAYERS I & II

Rank	1st	2nd	3rd	4th
Player I's Ranking	(5, 0)	(3, 3)	(1, 1)	(0, 5)
Player II's Ranking	(0, 5)	(3, 3)	(1, 1)	(5, 0)

Now, one could naturally raise the question as to which one of the four possible outcomes would be determined as the solution of the game. It is obvious from the table that this depends on not any one player's strategy choice but the combination of the two players' choices. Given the available options as shown in Table 4.1, each prisoner being rational in the sense of consistently seeking to maximize self-interests would obviously argue as follows:

Premise 1: Either the other player will cooperate with me or will defect against me.

Premise 2: If the other player cooperates with me, it is better for me to defect against than to cooperate with my opponent (because my defection against my opponent's cooperation yields 5 for me and 0 for him but my cooperation against his cooperation yields 3 for each of us).

Premise 3: If the other player defects against me, it is better for me to defect against than to cooperate with my opponent (because my defection against my opponent's defection yields 1 for each of us but my cooperation against his defection yields 0 for me and 5 for him).

Conclusion: Therefore, I am better off defecting (regardless of whether my opponent defects against me or not).

Thus each of the two players would choose the strategy of defecting regardless of whether the opponent chooses defection or cooperation. But a strategy that is the best for a player regardless of what the other player or players choose is a *dominant strategy*. If a rational player has a dominant strategy such that the remaining strategies are all dominated, then he or she would choose the dominant strategy and the dominated strategies will be eliminated. If there is a dominant strategy for each rational player, then each player will choose his or her dominant strategy and the *dominant strategy equilibrium* will be reached.

The upshot of the interactions, as shown by Table 4.2, is that both the players, guided by logic, defect against each other, and thereby each gets a payoff of 1, despite the possibility that had they cooperated with each other, each would have gained a higher

payoff of 3. We now try to shed light on this disappointing situation in terms of two extremely important solution concepts of game theory, viz., *Nash equilibrium* and *Pareto-efficiency* used to evaluate the different outcomes of a game. A *Nash equilibrium* is, as explained earlier, a profile of strategies, one for each player, such that each player's choice of strategy is the best reply given the other player's choice of strategy and hence no single player can obtain a higher pay-off by deviating unilaterally from this profile. An outcome or allocation of resources is, as explained earlier, *Pareto-efficient* (or, Pareto-optimal) if it is impossible to shift to any other outcome or reallocation of resources so as to make at least one individual better off without making any other individual worse off.

Now, let us consider the outcome (1,1) which, as shown by Table 4.2, is determined by mutual defection. The outcome (1,1) is a pure-strategy *Nash equilibrium*. Thus mutual defection is the uncoerced and *unique* Nash equilibrium and hence the only stable solution to PD. It may be pointed out that every dominant strategy equilibrium is by definition also Nash equilibrium, but every Nash equilibrium is not a dominant strategy equilibrium. Now the outcome (3,3) that would have resulted from mutual cooperation and is better for both the players than (1,1) is *not* a *Nash equilibrium*, because here it is *not true* that each player's choice of strategy is the best given the other player's choice of strategy just for the reason that a player can be better off by defecting if the other player cooperates. On the other hand, the outcome (3,3) is *Pareto-efficient*. But the outcome (1,1) is *Pareto-inefficient*, because it is an allocation of resources from which it is possible to shift to some other outcome or reallocation of resources, such as (3,3), so as to make at least one person better off without making any other person worse

off. In fact, a movement from outcome (1,1) to outcome (3,3) would be a shift from a *lose-lose* situation to a *win-win* situation that makes both persons better off.[16] The other two outcomes (5,0) and (0,5) are *Pareto-efficient* but not *Nash equilibrium*. Thus, the problem is that a group consisting entirely of individually rational decision makers collectively arrives at a situation that happens to be an *equilibrium* but *non-optimal* outcome.

It was stated earlier that Nash (1950, 1951) proved that every finite game has at least one Nash equilibrium, and that an equilibrium may be either in pure strategies or in mixed strategies. A *pure strategy* was defined as one that is used unconditionally, i.e., used with certainty or 100% probability. A *mixed strategy*, on the other hand, is one that assigns a positive probability to each pure strategy. Thus a pure strategy is an extreme instance of mixed strategy, because in the former the player assigns 100% probability to one strategy but only and 0% each of the remaining strategies. Since there is absolutely no reason for any player to play the strictly dominated strategy in any given play of the *Prisoner's Dilemma*, there is, by definition, no possibility of *mixed strategy equilibrium* in this game.

But Taylor (1976) and Axelrod (1981) have independently demonstrated that when there are repeated encounters in succession between the same two rational individuals facing the PD that satisfies the Ordering Condition as well as the Anti-exploitation Condition, the players get involved in what is known as an *Iterated Prisoner's Dilemma* (IPD) from which a very different situation may transpire and open

[16] It is important to understand that "both persons are better off" logically implies that "at least one person is better off"

new strategic possibilities to each player. First, each player in an IPD can *detect* the behavioral tendencies of his or her opponents from previous interactions. Secondly, as a consequence of the opportunity to know others, each player can *punish* by defection those who defected against cooperation but *reward* by cooperation those who cooperated against cooperation in the previous round so that for each player the best strategy to adopt is not defection but cooperation. Thirdly, in a repeated game each player gets an opportunity to develop a *reputation* for cooperation and thus to induce others to do the same. Consequently, in such circumstances a win-win situation through mutual cooperation may emerge as the *equilibrium* outcome (Hargreaves Heap & Varoufakis, 2004).

But a question arises as to whether the strategy of using *reputation* for cooperation to induce others to do the same will be sustainable and thereby lead to a Pareto-efficient as well as Nash equilibrium outcome. Obviously, repeated play of the game is necessary, because without it there can be no way to build reputation or to punish or reward the opponent. But then a question arises as to what sort of repetition is required? On this issue while the economist Hal Varian (2010) distinguishes between a *fixed* number of iterations and an *indefinite* number of iterations, the philosopher Steven Kuhn (2017) differentiates among *finite, infinite,* and *indefinite* number of iterations. Obviously, Kuhn's three-fold classification is exhaustive but overlapping because some finite iterations are indefinite when the players do not know exactly at which round the game will end and as such the future holds a prospect for both the parties. Moreover, infinite iteration may in a sense be considered as a subclass of indefinite iteration, because the future holds a prospect for the players as long as they gain a positive

marginal benefit from a repetition. Let us now briefly explain the concepts of finite, infinite, and indefinite iterations.

An IPD is *finite* if it involves a PD that is played by the same two players over exactly *n* successive rounds where n is a known counting number. Mutual defection is the unique Nash equilibrium in a finite IPD. This can be proved by means of the so called *backward induction* method that requires us to start at the last round and then keep going gradually backward from there to the first round. Consider, for example, a finite IPD of length n=5 in Table 4.4. In the last round the two players practically face a one shot PD with a dominant strategy and no one will have a chance to punish the opponent. So, both will defect in the last round. But if each player, being rational, knows that the other will cheat on the fifth and last round, then there will be no reason for anyone to cooperate in the penultimate or fourth round. The same reason applies till the first round. Therefore, both parties, as shown in Table 1.3, will defect throughout the game. An IPD is *infinite* if it involves a PD that is played by the same two players over *n* successive rounds where n tends to infinity. Mutual cooperation is obviously the Nash equilibrium in an infinite IPD, since each will have a chance to punish the opponent for there being no last round. The third and practically most interesting type of IPD is one of *indefinite* iteration which is stochastic by nature so that none of the players know the number of iterations in advance. Since in an indefinite IPD no one knows exactly when the game will terminate, there is no reason for any party to defect. Therefore, mutual cooperation will be the Nash equilibrium for an indefinite IPD.

TABLE 4.4: BACKWARD INDUCTION FOR A FINITE *IPD*					
Round No.	1	2	3	4	5
Player I's Move	D	D	D	D	D
Player II's Move	D	D	D	D	D

But a crucial question naturally arises as to whether there is any viable strategy that would ensure mutual cooperation as an optimal outcome in an IPD where different players could adopt different types of strategy. To address this question fully we need to clarify the meanings of some strategies used to play the IPD game. There are in fact many conceivable strategies for the IPD of which some of the simplest but most well-known ones are as follows: (i) All-C (Always Cooperate), (ii) All-D (Always Defect), (iii) RANDOM (Randomly Cooperate and Defect with equal probability), and (iv) TFT (Tit-for-Tat). TFT is defined as follows: Given any value of t where t stands for the number of rounds, cooperate in the first round, i.e., when t=1, and then on every subsequent round where t>1, do whatever your opponent did on the previous round, i.e., round t-1 (Axelrod, 1984).

The optimal strategy for a player in the one-time PD game is, as we have already seen, defection. But which strategy is optimal in an IPD game depends, as is evident from Table 4.2, on all the strategies in the environment. In an IPD game involving players who adopt All-D, the best strategy for a player is still All-D. When the opponents choose All-

C, the optimal strategy is again All-D. Moreover, when the opponents choose RANDOM, the best strategy is still All-D. And in a world where some contestants adopt All-D and some adopt All-C, the dominant strategy is yet again obviously All-D.

However, when TFT is introduced in a world populated by All-D, All-C, RANDOM, and various other strategies, what emerges as the best strategy is not necessarily All-D but rather depends on all the strategies in the population. In order to find optimal strategies Robert Axelrod conducted two computer IPD tournaments involving 14 strategies in the first and 63 in the second submitted by game theorists from several countries working in different disciplines (Axelrod, 1984). Surprisingly, in both the tournaments it was Tit-for-Tat (TFT) submitted by Russian-born Canadian mathematical psychologist Anatol Rapoport which became the winner by achieving the highest average score in spite of being the 'simplest' of all the strategies.

Axelrod (1984) succinctly stated his view about the reasons for the overall success of TFT over so many sophisticated strategies as follows:

> What accounts for TIT FOR TAT's robust success is its combination of being nice, retaliatory, forgiving, and clear. Its niceness prevents it from getting into unnecessary trouble. Its retaliation discourages the other side from persisting whenever defection is tried. Its forgiveness helps restore mutual cooperation. And its clarity makes it intelligible to the other player, thereby eliciting long-term cooperation (p. 54).

The *niceness* of TFT consists in never being the first to defect, and thus prevents TFT from getting into needless trouble. The *retaliation* of TFT signifies the property of

immediately punishing an unprovoked defection from the other side by defecting, and hence discourages the opponent from continuing defection whenever it is tried. The *forgiveness* of TFT refers to its propensity to forgive and forget any defection in the past which was already punished and to reward cooperation immediately by returning cooperation, and thereby helps restore mutual cooperation. The *clarity* of TFT indicates the property that makes it easy for the other player to recognize that the best way of coping with TFT is to enter into a relation of long-term mutual cooperation with it. Thus the reason why TFT succeeds is that it is a strategy of cooperation based on *niceness* as well as *reciprocity*. On the other hand, the first three strategies, All-C, All-D, and RANDOM, unlike TFT, prescribe a choice of action in advance, and hence cannot pay attention to knowing and thereby taking advantage of the opponent's moves in the previous rounds of the game before choosing an appropriate move in the current round.

Although TFT cannot beat any single contending strategy pitted against it, in Axelrod's tournaments it achieves remarkable overall success by surpassing all the rival strategies with respect to the cumulative total points scored. And the reason for TFT's robust success is due to its power of establishing cooperation. Even if TFT can create cooperation by virtue of being nice, retaliating, forgiving, clear, non-envious, transparent, and reciprocal, it ruthlessly punishes every defection and embodies the principle 'an eye for an eye and a tooth for a tooth'. However, after immediately punishing a defection TFT totally forgives and forgets that defection and hence exemplifies the principle 'let bygones be bygones'. Thus TFT has the advantage of reaping the full benefit of cooperation when matched against a cooperating opponent and avoiding the loss when matched against an unfriendly opponent. The combination of these two opposing

qualities of being nice and nasty results in a healthy policy expressed in the maxim 'live and let live' that makes it possible for cooperation to commence and continue.

Now, one might naturally ask the question "What, if any, is the best strategy for playing the IPD against a variety of strategies?" For quite a long time Axelrod's finding led most researchers to accept that TFT, while playing against various other strategies, would turn out to be the best or most successful strategy (Jurišić, Kermek, & Konecki, 2012). However, no strategy chosen by a player is the best independent of the strategy to be used by the opponent in a tournament, because the interests of the players in the PD are not in 'total conflict' and any one player's overall performance depends on the strategy choices of all the players (Axelrod, 1984; Fudenberg & Maskin, 1990; Nowak, 2006).

TFT has two weaknesses. First, if two TFT players are involved in repeated interactions with each other and one of them defects by an accidental mistake, then the other will retaliate, which in turn will lead to 'a sequence of alternating moves of cooperation and defection' that will result in disastrous consequences, i.e., very low payoffs for both of them, and there is no mechanism within the strategy for correcting the error by stopping unrelenting punishment and forgiving a single deviation in order to restore cooperation (Fudenberg and Maskin, 1990; Molander, 1985; Nowak and Sigmund, 1992). Second, suppose that there is a mixed population consisting of TFT and ALLC ('always-cooperate') players. In the absence of noise both strategies will continue to cooperate and will have the same average payoff. But the frequency of ALLC can rise due to random drift. When the number of unconditional cooperators exceed a certain threshold point players using ALLD ('always-defect') can invade the population. This

implies that TFT is *evolutionarily unstable* (Selten and Hammerstein, 1984; Boyd and Lorberbaum, 1987). Thus, although TFT was shown to be efficient in lots of IPD tournaments and was long considered to be the best strategy, it could be defeated in some specific circumstances (Wu and Axelrod, 1995; Beaufils, Delahaye, & Mathieu, 1996). Game theorists, therefore, have an enduring interest to find optimal or at least new strategies which would beat TFT in IPD tournaments.

4.3 The *Chicken*

The *Chicken* game, also known as the *Hawk-Dove* (HD) game, is, like the PD, a two-person, two-strategy, non-zero-sum, non-cooperative, mixed-motive, and ordinally symmetric game that captures some fundamentally important aspects of strategic interactions in conflict and cooperation in human and animal societies. The game of *Chicken* derives its name from a story that goes like this. Two people drive two fast cars towards each other from opposite ends of a long straight road. If one of them swerves before the other, he is called a chicken and the other driving straight a victor. Of course, if neither swerves, they will crash and meet the worst possible and disastrous consequence. Another possibility is that both drivers swerve. This is a case where neither has less honor than the other, and hence this is preferable to being a chicken. Thus the best outcome for each driver arises out of driving straight while the other one swerves. The second best outcome for each occurs when both swerve. The third best outcome for one is to swerve while the other drives straight. But the worst outcome for one is also the worst outcome for the other and happens when both drive straight. Although such games

can be played by immature, irresponsible, and reckless persons, politicians at the helm of state power have no right at all to play this sort of extremely risky game that would definitely jeopardize the life, property, and security of the innocent people. This is precisely the reason why Russell (1959) became so well-known for comparing the game of *Chicken* to nuclear brinkmanship which may lead to mutually assured destruction (MAD).

It may be pointed out that driving straight is comparable with the aggressive behavior of the hawk, while swerving with the peaceful behavior of the dove. Now, for each player let us assign payoff numbers to the four possible outcomes in accordance with the relative desirability of them as shown by the following payoff matrix:

TABLE 4.5: PAYOFF MATRIX FOR *CHICKEN* (or, *HAWK-DOVE*) *GAME*
(A Mixed-motive Game)

		Driver II	
Driver I	Strategy	Swerve (Dove): C	Drive Straight (Hawk): D
	Swerve (Dove): C	(R, R) (2, 2)	(S, T)* (1, 5)*
	Drive Straight (Hawk): D	(T, S)* (5, 1)*	(P, P) (-1, -1)

In the game of Chicken the cooperative strategy (C) based on the principle of 'live and let live' is obviously to swerve, and so the non-cooperative strategy (D) is to drive straight. Now, in terms of the four payoff numbers *P*, *R*, *S*, and *T* defined earlier as "Punishment for Mutual Defection", "Reward for Mutual Cooperation", "Sucker's Payoff for Cooperation against Defection", and "Temptation to Defect against Cooperation" respectively, we formally define the game of *Chicken* as one where each player ranks the outcomes in accordance with the following
preference relation:

T>R>S>P

Now, the game of *Chicken* is used to model many real life situations like labor-management conflict, international relations, and so on (Colman, 1999). The *Chicken* as a game-theoretic model is normally presented with informal narratives or stories supporting a typical social value orientation of players, viz., egoism. But the different games are defined and distinguished from one another in terms of not the informal narratives but rather the formal game structures or the preference orderings of the outcomes by the different players. For example, O'Henry's famous story of love and sacrifice *The Gift of the Magi* can be modelled as the game of *Chicken* (Anwar, 1999). To see how it can be done, let us make a short summary of the story as follows:

Della and her husband Jim were a happy couple who had two possessions in which they took a great pride. One was Jim's grand gold watch and the other was Della's

long and lustrous hair. They were both secretly planning to buy a special Christmas gift for one another. Unknown to Jim, Della sold her hair to buy a platinum fob chain to replace Jim's old leather strap of his gold watch. But unknown to Della, Jim sold his watch to get the money to buy for her a set of expensive combs that she longed for but could not buy. Thus, unfortunately, neither the chain could match the watch nor the combs the hair.

Obviously, there are two possible actions for each. Della could either ***sell*** or ***keep*** (i.e., not sell) her ***hair***. Similarly, Jim could either ***sell*** or ***keep*** his ***watch***. The two strategies of each of the two players would, therefore, jointly determine four possible outcomes – (Sell Hair, Sell Watch), (Sell Hair, Keep Watch), (Keep Hair, Sell Watch), and (Keep Hair, Keep Watch) – which may be symbolized respectively as (¬H, ¬W), (¬H, W), (H, ¬W), and (H, W). Now, the crucial question is: How would Della and Jim rank the four outcomes in accordance with their respective utility or preference functions?

We know that Both Della and Jim are *altruists*. So, each one would definitely like to make a sacrifice for buying a gift for the other. But that would definitely be incompatible with the altruistic goal of the other. As we can collect from O'Henry's story, each one has two conflicting desires – the desire to satisfy one's own unmet need and a much stronger desire to buy a gift for the other. Thus, Della would prefer sacrificing her beautiful hair for buying Jim a platinum fob chain for his gold watch on the assumption that all other things will remain the same. Similarly, Jim would prefer sacrificing his gold watch for buying Della an expensive comb for her lustrous hair on the assumption that all other things remain the same. Now, each of them assumes that all

other things will remain the same and secretly plans and proceeds to surprise the other with a special gift. But things do not always remain the same. And how things are depends not only on what someone is doing but also on what everyone else is doing. Thus genuine cooperation would require one to act towards helping the other person achieve her or his dominant goal.

Now, on the basis of the assumption of *non-satiety* which implies that human desires for wealth are never fully satisfied and so nobody would like to indulge in unnecessary waste of valuable resources and the assumption of *consistency* of human desires, it may be argued that Della's and Jim's rankings of the outcomes would reasonably be as shown in Table 4.6 (See Anwar, 1999 for detailed discussion):

TABLE 4.6: RANKINGS OF THE OUTCOMES BY DELLA & JIM

	1st	2nd	3rd	4th
Della	(¬H, W)	(H, W)	(H, ¬W)	(¬H, ¬W)
Jim	(H, ¬W)	(H, W)	(¬H, W)	(¬H, ¬W)

Though each of the agents is pursuing an altruistic goal, a *nontrivial game* situation arises. The reason for this is that though the two agents have altruistic goals, their goals are not entirely identical. As the table for the ordering of the outcomes shows, the second and the fourth choices of one player are

identical with the corresponding choices of the other player, but the first and the third choices of one are diametrically opposite to those of the other. As Harsanyi (1982/1990) puts it: "A nontrivial game situation can arise just as easily among altruists as it can among egoists - as long as these altruists are pursuing partly or wholly divergent altruistic goals (p. 43)."

Let us assume that the players can assign numerical values, such as 5, 2, 1, and -1, to the outcomes arranged in order of decreasing relative desirability of the outcomes for each of them. We can now build Table 4.7 showing the payoff matrix for the Della-Jim version of the game of *Chicken*:

TABLE 4.7: *CHICKEN* GAME (DELLA-JIM VERSION)

(A Mixed-motive Game)

	Jim		
	Strategy	Keep Watch (C)	Sell Watch (D)
Della	Keep Hair (C)	(R, R) (2, 2)	(S, T)* (1, 5)*
	Sell Hair (D)	(T, S)* (5, 1)*	(P, P) (-1, -1)

The justification for our modeling of the *Jim-Della game* as represented by the payoff matrix in Table 4.7 will be better understood if we compare it with a different matrix for the same situation modeled by Dixit and Nalebuff (1991, p.190) as follows:

TABLE 4.8: *DELLA-JIM GAME* (Dixit & Nalebuff, 1991, p. 190)

		Jim's Choice	
		Sell Watch (D)	Keep Watch (C)
Della's Choice	Keep Hair (C)	(1, 2)*	(0, 0)
	Sell Hair (D)	(0, 0)	(2, 1)*

The Dixit-Nalebuff matrix is similar to ours in so far as both matrices have two pure strategy Nash equilibria determined in each case by the same two strategy-combinations – (Keep Hair, Sell Watch) and (Sell Hair, Keep Watch). But Dixit and Nalebuff's matrix differs from ours in that they make no difference between the two outcomes (Sell Hair, Sell Watch) and (Keep Hair, Keep Watch) which they represent by the same pair of payoff numbers, (0, 0), whereas we make a clear distinction between the two as we represent the outcome (Sell Hair, Sell Watch) by the ordered payoff numbers (-1, -1) and (Keep Hair, Keep Watch) by (2, 2).

The reason why we differ with Dixit and Nalebuff is that their model fails to take account of the genuine difference between the two outcomes (Sell Hair, Sell Watch) and

(Keep Hair, Keep Watch). The outcome (Sell Hair, Sell Watch) involves a net loss of utility or benefit to the donors for ineffective sacrifice of two valuable things – the hair and the watch – without any corresponding gain of utility or benefit to the recipients. But the outcome (Keep Hair, Keep Watch) only maintains the *status quo* involving no loss and no gain and is even better than the wasteful sacrifices represented by the outcome (Sell Hair, Sell Watch).

There is a fundamental difference between our model represented by Table 4.7 and the Dixit-Nalebuff model denoted by Table 4.8, since the latter is presented rather as a coordination game only and, unlike the former does not satisfy the inequality condition "T>R>S>P" which is the formal definition of the game of *Chicken*. However, it may not go unnoticed that while Table 4.5 was backed by a narrative of egoistic or self-seeking players, Table 4.7 was supported by a narrative of altruistic or self-sacrificing players. Yet, Table 4.5 and Table 4.7 are fundamentally the same game, because both of them have two players, two strategies – cooperation and defection, exactly the same payoff numbers, and the same ordering of the outcomes by the players.

Let us now try to make a more detailed examination of why the sacrifices by the couple led to the unanticipated consequences. Altruism is generally acclaimed as an unconditional moral virtue, while egoism is commonly considered to be a vice. But, as pointed out earlier, uncoordinated and divergent goals of agents, however altruistic in spirit, will lead to a gaming situation involving conflict. Unplanned sacrifices by the agents are likely to be self-defeating and may end in repentance as well as rationalization of irrational act. This point has been aptly affirmed by Oakley, Knafo, and McGrath

(2011) as: 'What we value so much, the altruistic "good" side of human nature, can also have a dark side. Altruism can be the back door to hell (p. 4)'.

We now present two broad types of objections – a *logical* objection and a *psychological* objection – against wasteful and irrational altruism. What we call the *logical objection* concerns the logical absurdity of universal altruism. This point is put by Elster (1999), a notable scholar and proponent of analytical Marxism, as follows:

> … we cannot coherently imagine a world in which everyone had exclusively altruistic motivations. The goal of the altruist is to provide others with an occasion for selfish pleasures – the pleasure of reading a book or drinking a bottle of wine one has received as a gift. … If some are to be altruistic, others must be selfish, at least some of the time, but everybody *could* be selfish all the time (pp. 53-54).

The *psychological objection* against self-defeating altruism is based on a contrast between 'pathological altruism', a term coined by McWilliams (1984), and healthy altruism. Altruism as a 'caring concern' and an act for the wellbeing of others is definitely praiseworthy. But recent researches (McWilliams, 1984; Seelig & Rosof, 2001; Oakley, 2013; Sun, 2018; Kaufman & Jauk, 2020) have revealed that people in general are so blind to altruism that they fail to see that there may sometimes develop a dark side to it when an extremely strong altruistic motive of an individual to help others may turn into a hidden self-serving motivation so that the altruist ends up harming rather than helping others and yet is unable to become aware of the 'irrational illusions of helping'. This type of misplaced altruism is called *pathological altruism* and has been defined by Oakley, Knafo, and McGrath (2011) as follows:

> Pathological altruism might be thought of as any behavior or personal tendency in which either the stated aim or the implied motivation is to promote the welfare of another or others. But, instead of overall beneficial outcomes, the "altruism" instead has irrational and substantial negative consequences to the other or even to the self. (p. 4)

The corresponding definition of *pathological altruist* put forth by Oakley, Knafo, and McGrath (2011) is as follows:

> A person who sincerely engages in what he or she intends to be altruistic acts, but who harms the very person or group he or she is trying to help, often in unanticipated fashion; or harms others; or irrationally becomes a victim of his or her own altruistic actions. (p. 4)

From the above definition of pathological altruist displaying pathologically altruistic behavior, we can now bring out certain characteristics of a pathological altruist. First, the stated or the implied motivation of the pathological altruist is to promote the welfare of another or *others*. Second, the pathological altruist believes to be a sane and *rational* person pursuing the right ends with the right means. Third, such a misplaced altruist actually *deviates* from the original ends or means or both. Fourth, such an altruist is *unable reflect* on the original ends and means and so can utterly *lose control* of altruistic acts. Fifth, the pathological altruist actually creates *problems* not only for others but also

for himself or herself. In fact, some of world's most gruesome crimes, such as murder, genocide, suicide, and democrats turning into tyrant dictators can be explained to have been caused by misguided or pathological altruism. Hence, understanding the distinction between pathological altruism and healthy or genuine altruism is important for avoiding the detrimental consequences of the former and promoting the positive impacts of the latter.

Now, in the light of the above analysis Della's and Jim's altruism as wasteful, uncontrollable, and obsessive behavior can only be judged as irrational. In order to avoid settling for the worst and suboptimal outcome arising from mutual defection, (D, D) or (Sell Hair, Sell Watch), in this game, we now consider some possible optimal solutions of the Della-Jim version of the Chicken Game. Inspection of Table 4.5 and Table 4.7 and some calculation show that the game of chicken has ***two pure strategy Nash equilibria*** (D, C) and (C, D) and ***one mixed strategy*** (i.e., ***randomized***) ***Nash equilibrium*** where each player mixes the pure strategy of cooperation with probability 2/5 and the pure strategy of defection with probability 3/5. Further calculation shows that the expected payoffs of the two players in the mixed strategy Nash equilibrium are (7/5, 7/5) or (1.4, 1.4) in both the tables[17]. It is worth noting that these payoffs are fair but much lower than the best payoffs and slightly higher than the lower payoffs in the pure or deterministic Nash equilibrium outcomes, because randomization requires the agents to obtain even the worst outcome (D, D) some of the times.

[17] There is a standard procedure for calculation of the payoffs for a randomized Nash equilibrium which, of course, is quite complicated.

Let us now try to find the reason why the players fail to reach either one of the two mutually advantageous outcomes, (5, 1) or (1, 5). Can this failure be called a coordination failure that results from a failure of the players to coordinate their strategies so that they might settle for either (5, 1) or (1, 5)? To answer this question let us look at a semantic difference between 'coordination failure' and 'miscoordination' (Tyran, 2020). *Coordination failure* refers to an equilibrium phenomenon where a number of agents get stuck in a Pareto-inferior equilibrium outcome and fails to move to a Pareto-superior equilibrium outcome. *Miscoordination*, on the other hand, refers to a non-equilibrium phenomenon where a number of agents are trapped in an inefficient outcome and fail to move to a mutually advantageous outcome because of each player's uncertainty about the other player's move or lack of information about the gaming situation. In the light of this distinction we may now call the1 Della-Jim case a miscoordination rather than a coordination failure. Here the players fail to move not from an equilibrium but a non-equilibrium and inefficient outcome, (D, D) or (-1, -1), to an efficient as well as a mutually advantageous outcome , such as (5, 1) or (1, 5).

Let us now go back to an evaluation of the suboptimality of the mixed strategy Nash equilibrium that we mentioned earlier. The suboptimality of the mixed strategy equilibrium suggests that both players would do better if they could find a reliable way to *coordinate* their actions. Fortunately, we may find out a fourth equilibrium known as *correlated equilibrium* due to Nobel laureate game theorist Aumann (1974, 1987) which is a more general concept and computationally simpler than Nash equilibrium. A ***correlated equilibrium*** is an optimal outcome that can be obtained when players rationally coordinate their strategies through an assumed "trusted" authority that

randomizes among a selected number of outcomes with relatively high payoffs and tells each player what she or he is supposed to do and no player has an incentive to deviate from the suggested course of action.

Let us now see how to find a correlated equilibrium for the payoff matrix of *Chicken* as depicted in Table 4.7. Assume that there is a trusted third party that uniformly randomizes over the three outcomes (*C*, *C*), (*C*, *D*), and (*D*, *C*) on the basis of drawing one of three cards marked (*C*, *C*), (*C*, *D*), and (*D*, *C*) with probability l/3 each. The third party will then instruct each player to choose the strategy, C or D, required of her or him for producing the particular outcome shown by the draw. No player would like to deviate from the suggested course of action, since the expected payoff for deviating is lower than the payoff for obeying the instruction of the trusted party. It does not disclose to any player what the other player has been asked to do. Thus, the expected payoff for the correlated equilibrium is 2(1/3) + 1(1/3) + 5(1/3) = 8/3 = 2.67 (approx.) which is *fair* and *higher* than the expected payoff of the mixed strategy Nash equilibrium.

4.4 Similarities between the *Prisoner's Dilemma* and the *Chicken*

To grasp the similarities and the differences between the *Prisoner's Dilemma* and the *Chicken* it is helpful to begin with the formal definitions of the games, since most of their characteristics are inherent but implicit in, and so follows from, the definitions. Recall that the PD was defined by the *ordering condition* 'T>R>P>S', while the chicken by 'T>R>S>P'. But in case of repeated play of the PD another condition such as 'R>(S+T)/2' known as the *anti-exploitation condition* is added in order to disallow

players in taking turns to exploit each other by requiring the reward for mutual cooperation to be better than the average of the payoff for the temptation and the sucker. Let us now try to sort out and briefly state the similarities between these two important games.

First, the *Prisoner's Dilemma* and the *Chicken* are similar, since both of them are, as explained earlier, two-person, two-strategy, simultaneous-move, non-zero-sum, non-cooperative, mixed-motive, and ordinally symmetric games. Both of these are genuine games in which each agent ardently aims at attaining an articulated aspiration but gets involved in a strategic situation where what can be attained by anyone depends not only on what she or he does but also on what everybody else does.

Second, in both the PD and the *Chicken* the *best* or the most preferred payoff, from the individual's point of view, is D/C or T, since T>R, T>P, and T>S. This means that defection against cooperation is better than mutual cooperation, mutual defection, and cooperation against defection. Moreover, in both games R is a *fair* as well as the *second best* outcome, because it is better than both P and S, i.e., R>P and R>S. This means that the outcome from mutual cooperation, i.e., R is better than mutual defection and cooperation against defection. Yet, in neither of these games is the cooperative solution, i.e., (R, R), arising out of mutual cooperation is an equilibrium.

Third, both the PD game and the Chicken (HD) game involve a *social dilemma* since there is a conflict between individual rationality and collective rationality. Each of these games involves an *interpersonal* as well as an *intrapersonal* conflict. An interpersonal conflict refers to conflict between different persons over their beliefs,

values, interests, and actions, whereas an intrapersonal conflict refers to a conflict within a person arising out of choice over alternative beliefs, values, and actions.

The *interpersonal* conflict in the PD is evident from Table 4.2 which shows that the most preferred outcome for one player is the least preferred one for the other and vice versa, while the *intrapersonal* conflict consists in each player's vacillation between cooperation and defection when he or she expects the other player to cooperate. *Greed* or the desire to double-cross the opponent motivates one to defect against cooperation on the one hand, but *lack of trust* on the opponent's unenforceable agreement to cooperate induces one to defect against the expected cooperation on the other hand. Hence, both being equally rational decide to defect. Similarly, the interpersonal conflict in the HD game is evident from Table 4.7 which shows that the best or the most preferred outcome for one player is the third best preferred outcome for the other and vice versa, while the intrapersonal conflict consists in each player's hesitation about whether to cooperate or to defect depending on his or her expectation about the opponent's possible move.

Thus there is a vacillation in both the games. In case of the PD, it is between the Nash equilibrium outcome and the cooperative outcome that is better for both, Pareto-optimal, and fair. But in case of the *Chicken*, it is between P and T on the one hand and P and S on the other. In Chapter 2 a social dilemma has been as a situation where the individually reasonable behavior leads to an outcome in which everyone is worse off. By applying this definition it is easy to see that the payoff matrices of both the PD and Chicken satisfy the above definition of social dilemma.

Fourth, there is a *minimum safety level* in both the PD and the *Chicken*. In the PD the safety comes out of choosing *defection* (D) but in the Chicken it is secured from a choice of *cooperation* (C).

4.5 Differences between the *Prisoner's Dilemma* and the *Chicken*

As pointed out earlier the differences between the *Prisoner's Dilemma* and the *Chicken*, like the similarities between them, arises mainly from their formal definitions. We now point out the main differences between the two games as follows:

First, the most fundamental difference between the two games is that in the PD we have P>S, but in the *Chicken* S>P. This means that in the PD not cooperating with the non-cooperator is better than cooperating with her or him, while in the *Chicken* cooperating with the non-cooperator is better than not cooperating with her or him. The reason for this is quite transparent. In the *Chicken* the worst or disastrous outcome is P which occurs if both players defect, whereas the worst outcome in the *Prisoner's Dilemma* is S which happens to one who cooperates with a defector. It may be pointed out that our feelings about the type and magnitude of the worst outcome is a psychological matter that arises out of the story associated with a model and is not truly reflected in the payoff numbers.

Second, the minimum safety level in the PD and the *Chicken* are different. In the PD it arises out of choosing defection. And all the players tend to choose defection. As a result the outcome (D, D) is determined as the equilibrium. But in the Chicken the minimum safety level is assured when at least one player chooses cooperation. But even if both the players choose cooperation, the outcome determined (C, C) is not in equilibrium but yet Pareto optimal.

Third, in the PD each player has a *dominant strategy* that leads to a *dominant strategy equilibrium*, while in the *Chicken* no player has a dominant strategy. This dominant strategy equilibrium is also the only *pure strategy Nash equilibrium* in the PD but is *Pareto-inefficient*. There is no *mixed strategy equilibrium* and no *correlated equilibrium* in the PD, since the dominant strategy is chosen with 100% probability. In the *Chicken* no player has a dominant strategy, and so there is no dominant strategy equilibrium. But in this game there are two pure strategy Nash Equilibria that happen to be Pareto-optimal but asymmetric and unfair. In addition to these, there are two more Equilibria in the *Chicken* –a *mixed strategy Nash equilibrium* and a *correlated equilibrium* that is fair and Pareto-optimal.

CHAPTER FIVE

EGOISTIC COOPERATION AND ALTRUISTIC DEFECTION: INDIRECT PROOFS

5.1 Preview

In chapter 1 we discussed five possible responses to our research question about the nature of the causal connection between altruism and cooperation. In this chapter we formulate two different arguments and prove their validity in order to refute the third of those responses which happens to be the prevailing view or hypothesis that altruism is

either a necessary or a sufficient condition for cooperation to occur. One argument is intended to show that altruism is not necessary for cooperation, that is, non-altruists or egoists can cooperate. The other argument is intended to show that altruism is not sufficient for cooperation, that is, altruists may fail to cooperate. But before we can construct our arguments, we need to sort out the relevant facts and conjectures about the relation between altruism and cooperation. After showing that altruism is neither necessary nor sufficient for cooperation, we explain the view that altruism still an important "contributory cause" of cooperation.

5.2 Facts and Conjectures about Cooperation

As we want to nullify the view that altruism is either necessary or sufficient for cooperation, we treat it as the null hypothesis (H_0). We, therefore, want to prove the alternate hypothesis (H_a) that altruism is *neither* a necessary *nor* a sufficient condition for cooperation which is the negation of the null hypothesis. We now state, rephrase, and symbolize these two hypotheses as follows:

(1) "Altruism is either a necessary or a sufficient condition for cooperation to occur". This may be equivalently rephrased as "All cooperators are altruists or all altruists are cooperators". This statement may now be translated into symbolic language as:

$(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)$

(2) "Altruism is neither a necessary nor a sufficient condition for cooperation" This may be equivalently paraphrased as "It is false that either all cooperators are altruists or all altruists are cooperators". We now translate this statement into symbolic language as:

$\neg[(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)]$

Now, the null hypothesis and the alternate hypothesis will be as follows:

Null Hypothesis (H_0): $(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)$

Alternate Hypothesis (H_a): $\neg[(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)]$

Now that we have the relevant hypotheses about causal relationships between altruism and cooperation, we would require sorting out and adding to these the appropriate empirical facts about altruism and cooperation. Analysis of two important game-theoretic models in chapter 4 and our review of literature in chapter 2 have firmly established certain empirical facts about the agents being either egoistic or altruistic in attitude and being cooperative or non-cooperative in acts under strategic situations. For example, the Prisoner's Dilemma model, as discussed in chapter 4, shows that when rational egoistic agents pursuing their own interests have an opportunity to interact again and again for an unknown or indefinite (not necessarily infinite) number of rounds, mutual cooperation frequently emerges and develops among them (Axelrod, 1984). On the other hand,

our introspection and thought experiment on O'Henry's famous story of love and sacrifice "*The Gift of the Magi*" modeled as the Chicken (Anwar, 1999), as discussed in chapter 4, shows that when each of two altruistic agents is independently pursuing the interests of the other in a strategic environment, they may arrive at the worst possible and non-cooperative outcome.

Now, if we sort out the information mentioned above, we may state it in just two existential statements: (a) "There are egoists who cooperate". This may be rephrased as "Some egoists are cooperators". (b) There are altruists who fail to cooperate. This may be rephrased as "Some altruists are non-cooperators". Now these two statements may be symbolized as follows:

(a) $(\exists x)(Ex \wedge Cx)$

(b) $(\exists x)(Ax \wedge \neg Cx)$

We now need one more thing to do in order to construct our arguments. We have to relate the two terms "altruist" and "egoist", because formal logic would not know the semantic difference between them. Since our research question was, in conformity with the dissertation title, framed in terms of the word "altruism" and we regard "altruism" and "egoism" as contradictory terms, we would *define* the term "egoist" as "non-altruist" and convert the singular statement "x is an egoist" into "x is not an altruist", i.e., "Ex" into "$\neg$Ax".

Now, to nullify the null hypothesis "$(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)$", we have to show that "$(\forall x)(Cx \rightarrow Ax)$" is false and "$(\forall x)(Ax \rightarrow Cx)$" is false. That is, we have to prove that "$\neg(\forall x)(Cx \rightarrow Ax)$" is true and "$\neg(\forall x)(Ax \rightarrow Cx)$" is true.

5.3 Argument for Egoistic Cooperation

The argument for egoistic cooperation is intended to show that "$(\forall x)(Cx \rightarrow Ax)$" is false, i.e., altruism is not necessary for cooperation. But the null hypothesis assumes that this is true. Now, if accepting this as true along with the empirical fact "$(\exists x)(Ex \wedge Cx)$" and the definition "$Ex=\neg Ax$" leads to an explicit contradiction, then the assumption *cannot* be true, and hence must be false. Before we proceed to construct an indirect proof of this claim, let us put together the facts, conjectures, and definition as shown below:

Null Hypothesis (H_0): $(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)$

Alternate Hypothesis (H_a): $\neg[(\forall x)(Cx \rightarrow Ax) \vee (\forall x)(Ax \rightarrow Cx)]$

Definition: $Ex = \neg Ax$

Empirical Fact (a): $(\exists x)(Ex \wedge Cx) = (\exists x)(\neg Ax \wedge Cx)$ [By Definition]

Empirical Fact (b): $(\exists x)(Ax \wedge \neg Cx)$

Now, we can build up the *indirect proof* (See Hurley, 2012 for a description) as follows:

1.	$(\exists x)(Ex \wedge Cx)$	Given Empirical Fact
2.	$(\forall x)(Cx \rightarrow Ax)$	Assumption for Indirect proof
3.	$(\exists x)\ (\neg Ax \wedge Cx)$	1, Definition of Ex as $\neg Ax$
4.	$\neg Aa \wedge Ca$	3, Existential Instantiation
5.	$Ca \rightarrow Aa$	2, Universal Instantiation
6.	$\neg Aa$	4, Simplification
7.	$\neg Ca$	5, 6, Modus Tollens
8.	$Ca \wedge \neg Aa$	4, Commutation
9.	Ca	8, Simplification
10.	$Ca \wedge \neg Ca$	9, 7, Conjunction
11.	$\neg(\forall x)(Cx \rightarrow Ax)$	2 – 10, Indirect Proof

The method of indirect proof used above was explained in chapter 3. The above proof for the argument consists of eleven statements numbered sequentially from (1) to (11) where the justification for each statement in the proof is written on the right side of it. Here statement (1) is an empirically established fact serving as the premise of the argument. Statement (2) is the one that we want to show is false and hence has been assumed as true to serve as an *assumption for indirect proof* (AIP). The indirect proof

sequence begins with the AIP at line (2) and ends at line (10) where the AIP leads to an explicit contradiction that entitles us assert the denial of the AIP as the conclusion in line (11).

5.4 Argument for Altruistic Defection

The argument for altruistic defection is intended to show that "$(\forall x)(Ax \rightarrow Cx)$" is false, i.e., altruism is not sufficient for cooperation. But the null hypothesis assumes that this is true. Now, if accepting this as true along with the empirical fact "$(\exists x)(Ax \wedge \neg Cx)$" leads to an explicit contradiction, then the assumption *cannot* be true, and hence must be false.

Below we now construct an *indirect proof* for the claim that "$(\forall x)(Ax \rightarrow Cx)$" cannot be true, and hence must be false as follows:

1. $(\exists x)(Ax \wedge \neg Cx)$ Given Empirical Fact
2. $(\forall x)(Ax \rightarrow Cx)$ Assumption for Indirect proof
3. $Aa \wedge \neg Ca$ 1, Existential Instantiation
4. $Aa \rightarrow Ca$ 2, Universal Instantiation
5. Aa 3, Simplification
6. Ca 4, 5, Modus Ponens
7. $\neg Ca \wedge Aa$ 3, Commutation
8. $\neg Ca$ 7, Simplification
9. $\neg Ca \wedge Ca$ 8, 6, Conjunction
10. $\neg(\forall x)(Ax \rightarrow Cx)$ 2 – 9, Indirect Proof

The above proof for the argument consists of ten statements numbered sequentially from (1) to (10) where the justification for each statement in the proof is written on the right side of it. Here statement (1) is an empirically established fact used as the premise of the argument. Statement (2) is what we want to show false and hence has been assumed as true to serve as an *assumption for indirect proof* (AIP). The indirect proof sequence begins with the AIP at line (2) and ends at line (9) where the AIP leads to an explicit contradiction that permits us assert the denial of the AIP as the conclusion in line (10).

5.5 Altruism as a Contributing Condition for Cooperation

Before we begin our discussion of whether altruism can at all contribute towards creating a cooperative culture, we will consider the implication of the two arguments used to refute the null hypothesis that cooperation is either a necessary or a sufficient condition for cooperation. Have we unmistakably refuted this hypothesis? Logical analysis shows that the null hypothesis being a disjunction of two universally quantified statements can never be conclusively verified or confirmed. Moreover, Karl Popper (1959, 1962), the most highly respected, and perhaps the greatest, philosopher of science of the twentieth century, holds that the true purpose of science, as opposed to pseudo-science, is not to verify or confirm general scientific theories or conjectures by empirical evidence but rather to refute or falsify them by counterexamples obtained from rigorous

tests. Now, our alternate hypothesis being the negation of the disjunction of two universally quantified statements turns out on analysis to be the conjunction of two existentially quantified statements. Since the null hypothesis is a non-probabilistic and universal statement and has been shown to be in direct conflict with our empirical facts, it may be treated as conclusively falsified leaving no room for mistakenly rejecting it when it is true or accepting it when it is false. But this means that altruism is in fact neither necessary nor sufficient for cooperation.

Thus the only possibility of explaining any causal relation between altruism and cooperation that remains open to us is the fifth sense. Thus while scholars in general (such as, Hatcher et al., 1990; Dissanayake & Bezwada, 2010; Copi, Cohen, & McMahon, 2014, Kelley, 2014) have begun to appreciate the necessity of a relatively "weaker" concept of cause as unnecessary and insufficient but still 'contributory' condition, epidemiologists in particular (such as, Riegelman, 1979, 2012) have begun to develop the idea of cause as a 'contributory' cause for non-communicable diseases, death, and accident. For example, lung cancer that can occur in the absence and fail to occur in the presence of smoking is both unnecessary and insufficient but still a 'contributory' condition that can in conjunction with other relevant factors cause the disease. Traditional medical science, such as Koch's postulates, once used to list a set of necessary and sufficient conditions for bacterial diseases (Riegelman, 2012). But the possibility of identifying all the necessary conditions that together constitute the sufficient condition of a disease has been challenged on the ground of the uncertainty as to whether germs cause a disease or the disease causes the germ.

Riegelman (2012) has offered a definition of 'contributory cause' that must satisfy all of the following three conditions: (1) The "cause" and the "effect" are associated, i.e., occur together, at the individual level. (2) The "cause" precedes the "effect" in time; and (3) altering the "cause" alters the probability of the "effect". The presence or absence of the contributory cause does *not ensure* but does *contribute* to the presence or absence of the effect. While its presence increases the probability of occurrence of the effect, its absence reduces the probability of the effect. There may be multiple contributory causes that may often work jointly to produce the same effect. Multiple interventions may modify the cause, and may thereby modify the effect.

Thus, a possible explanation of why the two terms "altruism" and "cooperation" are often used synonymously is that people find it difficult to disentangle them because of the widespread conviction that agents with altruistic attitude are obviously more likely to cooperate than not. Altruism is unnecessary and insufficient but still a contributing factor for cooperation. There is a need for further research to find out: (1) how to alter people's attitude to make them contribute more towards creating a cooperative culture, (2) whether other conditions such as agents' age, sex, profession, contact, or cultural factors can work as contributory causes, (3) whether altruism in conjunction with some other conditions can work better to lead to cooperation, (4) how to intervene to reduce or eliminate cooperation among agents engaged in organized crimes.

CHAPTER SIX

CONCLUSION

6.1 Preview

This chapter has a three-fold purpose. These are: (i) To provide a *brief summary* of the previous chapters based on the central research question and other related questions; (ii) To recapitulate some of the main research *findings* of the study; and (iii) To indicate some places in the dissertation where we could offer *suggestions for further research*.

6.2 Brief Summary of the Study

This dissertation has been divided into six chapters each of which includes the discussion of a particular topic and the conclusion that follows. In this section I present a brief summary of the main topics discussed in each chapter.

In Chapter 1 I have formulated the main research problem, discussed its background, indicated the justification for doing this research, and finally outlined the organization of the dissertation.

The *central research problem* formulated as a lengthy and compound question is as follows: Is altruism only a necessary condition, or else only a sufficient condition, or else either a necessary or a sufficient condition, or else both a necessary and a sufficient condition, or else neither a necessary nor a sufficient condition for mutual cooperation among rational individuals interacting in circumstances where their interests are neither entirely identical nor completely contradictory, but rather mixed that leave open the possibility of cooperation as well as defection?

To fully answer the above question, an additional research question has been raised. Having argued that the right answer to the main question is that altruism is neither a necessary nor a sufficient condition for mutual cooperation to happen under the stated circumstances, we have been naturally led to a second but important question as follows: If altruism is neither a necessary nor a sufficient condition, can it still be a ‘contributory condition’ for cooperation?

Chapter 2 has provided a comprehensive review of the relevant literature and conceptual framework of the study. In Section 2.2 I have briefly explained and evaluated the rational choice theory which involves methodological individualism as opposed to methodological collectivism and seeks to establish macro-behavior of humans on a micro-foundation. In Section 2.3 I have discussed the problem of rationality which is related to the rational choice theory and deals with whether humans are rational or not. Subsections 2.3.1, 2.3.2, and 2.3.3 have comprehensively examined the various meanings and types of rationality, the nature and conditions of instrumental rationality, and the limits of instrumental rationality, respectively. Section 2.4 has been divided into two subsections of which subsection 2.4.1 has presented a detailed discussion of the elements of game theory and subsection 2.4.2 has discussed the distinction among private, public, common, and club goods, and the nature and types of social dilemma games particularly the common good dilemma and the public good dilemma. Section 2.5 has been divided into three subsections and has dealt with Darwin's theory of natural selection, its limits, and the Darwinian puzzle as a background to the theories of cooperation. Section 2.6 has been divided into seven subsections and each subsection has dealt with one of the seven different mechanisms of cooperation, viz., Kin Selection, Group (or Multilevel) Selection, Spatial Selection, Direct Reciprocity, Indirect Reciprocity, Strong Reciprocity, and Costly Signaling. Section 2.7 has examined the Egoism-Altruism Debate. Section 2.8 has been divided into two subsections of which subsection 2.8.1 has discussed the concept of conflict and its various types such as interpersonal, intrapersonal, and group conflicts and subsection 2.8.2 has discussed the nature and types of cooperation.

Chapter 3 has been devoted to the discussion of methodological issues. Subsection 3.2 has distinguished between proof and disproof and then between two types of proof – direct proof and indirect proof. Subsection 3.3 has distinguished among the three related concepts of conditions, causation, and correlation in order to shed light on the concept of cause.

In Chapter 4 I have discussed the *Prisoner's Dilemma* and the *Chicken* (or *Hawk-Dove Game*) as two different mixed-motive models of strategic interactions. In Subsection 4.2 I have thoroughly examined the one-shot *Prisoner's Dilemma* as a matrix (or strategic) form game, the *Iterated Prisoner's Dilemma* (IPD), and several strategies such as, All-D, All-C, Random, and TFT, for dealing with numerous real life situations of conflict and cooperation. In Subsection 4.2 I have examined the Chicken (or Hawk-Dove) game and have presented it first with the usual example of egoistic players and then with a novel example of altruistic players based on O'Henry's celebrated story of love and sacrifice *The Gift of the Magi*. In Subsections 4.4 and 4.5 I have discussed the similarities and the differences between the two games in the light of the concepts of *Dominant Strategy Equilibrium*, *Nash Equilibrium*, *Correlated Equilibrium*, and *Pareto Optimality*.

In Chapter 5 this dissertation may be said to have reached its culmination in the sense that the *indirect proofs of validity* for two arguments of which one refutes the view that altruism is a necessary condition for cooperation and the other refutes the view that altruism is a sufficient condition for cooperation. Then we have argued that just as smoking is neither a necessary condition nor a sufficient condition but is nevertheless a contributing condition for lung cancer, so also altruism is neither a necessary condition nor a sufficient condition but is yet a contributing condition for cooperation.

6.3 Some Research Findings

In this section we put together not all but only some of findings of the study collected from the previous chapters.

First, one important finding from chapter 2 is about Aristotle's definition of man as a rational being encapsulated in the maxim "Man is a rational animal" and Russell's humorous comment to it for not being able to find even a single positive instance of rational man in spite of his serious search for one over his long life. But we have found that this objection does not succeed as it is based on an *equivocation* between defining *mankind* or a *potential man* as rational and describing an actual *individual man* who may fall short of the potentially rational man. Potentiality should not be tied down or equated to any particular individual man. *Aristotle's maxim*, therefore, *still stands*.

Second, as to the question whether the traditional *consistency condition of rationality* which is extremely unrealistic and unattainable for boundedly rational humans should be rejected, replaced, or retained, Sen (1990) argued against the rejection because of the difficulty of finding an alternative set of simpler assumptions. But I argued that it is more reasonable to retain than to reject the traditional consistency condition of rationality not because there is not yet any better alternative but because we do always and necessarily *need an ideal which must have a gap* so that it allows us to make improvement of practice.

Third, the *egoism-altruism dichotomy* discussed in chapter 2 is *unsustainable*, because neither universal egoism nor universal altruism is defensible either as an empirical or a normative principle. People can behave egoistically at some times and

altruistically at other times. As a matter of fact, even standard economic models of rational behavior do not presuppose that men are essentially selfish in the sense in which it is interpreted in the egoism-altruism debate.

Fourth, we must guard against a *dogmatic assumption* that even educated laymen usually make about Darwin's theory of natural selection and the survival of the fittest. Darwin's concept of the survival of the fittest discussed in chapter 2 does not, as presumed by many people, mean that to be fit is to be physically stronger or mentally sharper than others who are not fit. Fitness only refers to an attribute which is favored by nature.

Fifth, it is important to note that *caution* should be exercised in the empirical study of games. Although formal game theory is an intellectually rewarding and practically useful tool, we must be careful about drawing any conclusion from an empirical investigation of games in real life. Games are mathematical models that are defined only in terms of the relative sizes of payoff numbers of the players involved and are completely devoid of the any psychological or ideological conditions of them that are not taken account of by the numbers. The same formal game can be used to illustrate different situations as long as the relative rankings of the outcomes by the players are the same. For example, in chapter 4 the game of Chicken was reasonably used to illustrate two different situations in one of which the players were egoistic in nature while in the other the players were altruistic.

Sixth, we can learn an extremely important *lesson* about how to manage an organization efficiently and effectively from the concept of *correlated equilibrium* invented

by Nobel laureate game theorist Aumann (1974, 1987) which is a more general concept and computationally simpler than Nash equilibrium. A correlated equilibrium, as defined earlier, is an optimal outcome that can be achieved when players rationally coordinate their strategies through an assumed "trusted" authority that randomizes among a selected number of outcomes with relatively high payoffs and tells each player what she or he is supposed to do and no player has an incentive to deviate from the suggested course of action. An organization consisting of members with partially conflicting interests can attain a higher as well as fair outcome for all, provided there exists an authority that is "trusted" and trustworthy.

Seventh, an important general lesson that emerges from the study is that isolated, independent, and individualistic decision, whether motivated by a self-seeking or a self-sacrificing impulse, has to be prevented in order to avoid unnecessary mutual loss. It has been shown that secret and *uncoordinated* attempts at being altruistic toward others may lead to a shock from unanticipated loss rather than a surprise from expected gain for all the agents in a group.

Eighth, one of the most important results shown by this dissertation is the construction two arguments that have been proved to be logically valid by the method of indirect proof and shown to be sound as well. These two arguments known as the Argument for Egoistic Cooperation and the Argument for Altruistic Defection are used to prove two counterintuitive truths that cooperation can take place among egoists and that cooperation may fail to take place among altruists, respectively.

6.4 Suggestion for Further Research

There are two important aspects of the dissertation where further research could be done to extend our knowledge. One is to explore the concept of "contributory cause" as it applies particularly to the causal connection that exists between "altruism" and "cooperation" when altruism is unnecessary and insufficient for producing cooperation. The other possible aspect for extending the research would be to explore the causal link between altruism and cooperation in terms of the concept of counterfactual conditionals.

BIBLIOGRAPHY

Alexander, R. D. (1987). The biology of moral systems. New York: Aldine de Gruyter.

Ale, S.B., Brown, J.S., Sullivan, A.T. (2013). "Evolution of Cooperation: Combining Kin Selection and Reciprocal Altruism into Matrix Games with Social Dilemmas". PLoS ONE 8(5): e63761. doi:10.1371/journal.pone.0063761.

Al-Ghazali, A.H.M. (2000). *The Incoherence of the Philosophers*, M. E. Marmura (trans.), Provo, Utah: Brigham Young University Press.

Andreoni, J. and J. Miller (1993) 'Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence'. *Economic Journal*, 103, 570-85.

Anwar, A.J. (1996). "Resolving the Paradox of Confirmation: A Study in Material Implication and Information". In: *Chittagong University Studies* (Arts), Vol. XII, pp. 187-204.

Anwar, A.J. (1999). "Modelling Altruistic Behaviour: A Case of Failure in Coordination". In: *Humanomics*, Vol. 15, Issue: 4, pp. 94-122, http://dx.doi.org/10.1108/eb018841

Ardrey, R. (1970). *The Social Contract*. London: Collins.

Aristotle. (1941). *Nicomachean Ethics*, trans. W.D. Ross. Random House, New York. (Bk.1, Ch.13)

Arrow, K.J. (1963). *Social Choice and Individual Values*, 2nd Ed., John Wiley, New York.

Aumann, R.J. (1974). "Subjectivity and Correlation in Randomized Strategies". In: *Journal of Mathematical Economics*. Vol. 1, pp.67-96.

Aumann, R.J. (1987). "Correlated Equilibrium as an Expression of Bayesian Rationality". In: *Econometrica*, Vol. 55, No. 1, pp. 1-18.

Axelrod, R. (1970). *Conflict of Interest: A Theory of Divergent Goals with Applications to Politics*, Markham Publishing Co., Chicago.

Axelrod, R. (1981). "The Emergence of Cooperation among Egoists", *American Political Science Review*, Vol. 75, No. 2, pp. 306-318.

Axelrod, R., & Hamilton, W.D. (1981). The Evolution of Cooperation. *Science*, Vol. 211, No. 4489, pp. 1390-1396. Stable URL: http://www.jstor.org/stable/1685895.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, New York.

Axelrod, R. (1997). *The Complexity of Cooperation*. Princeton Univ. Press, Princeton, NJ.

Axelrod, R., & Cohen, M.D. (2000). *Harnessing Complexity: Organizational Implications of a Scientific Frontier*, Free Press, New York.

Ball, M.A. (1985). *Mathematics in the Social and Life Sciences*. West Sussex: Ellis Horwood.

Bardis, P.D. (1979). "Social Interaction and Social Processes", *Social Science*, Vol. 54, No. 3, pp. 147-167. <http://www.jstor.org/stable/41886414>

Baron, J. (2008). *Thinking and Deciding*, 4th Ed., Cambridge University Press, New York.

Barros, G. (2010). "Herbert A. Simon and the Concept of Rationality: Boundaries and Procedures". In *Revista de Economia Política*, Vo. 30, pp. 455-472.

Batson, C.D. & Shaw, L.L. (1991). "Evidence for Altruism: Toward a Pluralism of Prosocial Motives". In: *Psychological Inquiry*, Vol. 2, No. 2, pp. 107-122. Stable URL: http://www.jstor.com/stable/1449242

Batson, C. D. (2010). "Empathy-induced Altruistic Motivation". In: M. Mikulincer & P. R. Shaver (Eds.), *Prosocial Motives, Emotions, and Behavior: The Better Angels of Our Nature* (p. 15–34). American Psychological Association. https://doi.org/10.1037/12061-001

Batson, C. D. (2011). *Altruism in Humans*. Oxford University Press, New York.

Batson, C. D. (2014). *The Altruism Question: Toward a Social-Psychological Answer*. Psychology Press, New York.

Bardis, P.D. (1979). "Social Interaction and Social Processes", *Social Science*, Vol. 54, No. 3, pp. 147-167. <http://www.jstor.org/stable/41886414>

Baumol, W.J. (1982). *Economic Theory and Operations Analysis*, 4th Ed., Prentice-Hall, New Delhi.

Baurmann, M. (1998). "Liberal Society and Planned Morality?" in E. Morscher et al. (eds.), *Applied Ethics in a Troubled World*, pp. 203-223, Kluwer Academic Publishers, Netherlands.

Beaufils, B., Delahaye, J., & Mathieu, P. (1996). Our meeting with gradual: A good strategy for the iterated prisoner's dilemma, *Proceedings of the Artificial Life V*, pp. 202-209.

Becker, G. (1974). "A theory of social interactions," *Journal of Political Economy*, 82(6):1063–93.

Becker, G.S. & Murphy, K.M. (1988). "A Theory of Rational Addiction", *The Journal of Political Economy*, Vol. 96, No. 4, pp. 675-700.

Beggs, J. (2020, February 11). *Positive versus Normative Analysis in Economics*. Retrieved from https://www.thoughtco.com/positive-versus-normative-analysis-1147005

Bell, D.E., Raiffa, H., & Tversky, A. (1988). "Descriptive, Normative, and Prescriptive Interactions in Decision Making", in *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Bell, D.E., Raiffa, H., & Tversky, A. (Eds.), Cambridge University Press, New York. (Ch. 1, Pp. 9-30)

Bell, D.E., Raiffa, H., & Tversky, A. (1988). "Descriptive, Normative, and Prescriptive Interactions in Decision Making", in *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Bell, D.E., Raiffa, H., & Tversky, A. (Eds.). NY: Cambridge University Press. (Ch. 1, Pp. 9-30)

Bennett, D.J. (2004). *LOGIC MADE EASY*. Norton & Co., Inc., New York, N.Y.

Bergin, J. (2005). *Microeconomic Theory: A Concise Course*. New York: Oxford University Press.

Binmore, K. (1992) *Fun and Games: A Text on Game Theory*, D.C. Heath, Lexington, MA.

Binmore, K. (2004). "Reciprocity and the Social Contract". In *Politics, Philosophy & Economics*, Vol 3, Issue 1, pp. 5 – 35. https://doi.org/10.1177/1470594X04039981

Bowles, S., Fehr, E., and Gintis, H. "(2003) 'Strong Reciprocity and the Conditions for the Evolution of Altruism" [Internet paper].Google Scholar.

Braithwaite, R. B. (1955). *Theory of Games as a Tool for the Moral Philosopher*, Harvard University Press, Cambridge, MA.

Broad, C. D. (1950). "Egoism as a Theory of Human Motives", *Hibbert Journal* 48:105-114.

Brossel, P., Eder, A., & Huber, F. (2013). Evidential Support and Instrumental Rationality. In *Philosophy and Phenomenological Research*. Wiley Online, Vol. LXXXVII, No. 2, doi- 10.1111/j.1933-1592.2011.00543.

Brue, S.L., McConnell, C.R., & Flynn, S.M. (2014). *Essentials of Economics*, 3e, McGraw-Hill, New York.

Buchanan, J. M., & Tullock, G. (1962). *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.

Butler, J. (1729). *Fifteen Sermons Preached at Rolls Chapel*, 2nd ed. J. and J. Knapton, London.

Cahn, S.M. (2013). "The Altruism Puzzle", *J. of Soc. Philos*, 44 (2): p. 107. doi:10.1111/josp.12023.

Campbell, R. & Sowden, L. (Eds.). (1985). *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, the University of British Columbia Press, Vancouver, 1985.

Capraro, V. (2013). "A Model of Human Cooperation in Social Dilemmas". *PLoS ONE*, Vol. 8, Issue 8, pp. 1-6. e72427. doi:10.1371/journal.pone.0072427.

Carballo, D. M. (2012). "Cultural and Evolutionary Dynamics of Cooperation in Archaeological Perspective", D. M. Carballo (Ed.), *Cooperation and Collective Action: Archaeological Perspectives*. Boulder: University Press of Colorado.

Carey, Gregory. (2003). "Chapter 13: The Five Forces Behind Human Evolution", *Human Genetics for the Social Sciences*, SAGE Publications, Inc. DOI: http://dx.doi.org/10.4135/9781452229591.n13. Access Date: Nov. 27, 2017.

Carroll, L. (1865/2018). *Alice's Adventures in Wonderland*, Global Grey, globalgreyebooks.com

Carter, T. (2007). "Public Goods Dilemma". In: R.F. Baumeister, & K.D. Vohs, (Eds.), *Encyclopedia of Social Psychology*, SAGE Publications, Inc., Thousand Oaks. http://dx.doi.org/10.4135/9781412956253.n428. Access Date: September 24, 2019.

Cato, S. (2013). "Social Choice, the Strong Pareto Principle, and Conditional Decisiveness". In: *Theory and Decision*, **75,** 563–579. https://doi.org/10.1007/s11238-013-9352-9

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Clavien, C. (2012). "Altruistic Emotional Motivation: An Argument in Favour of Psychological Altruism". In: K. Plaisance & T. Reydon (Eds.), *Philosophy of Behavioral Biology*. Boston Studies in Philosophy of Science, Springer Press.

Clavien, C. & Chapuisat, M. (2013). "Altruism across Disciplines: One Word, Multiple Meanings", *Biology and Philosophy* Vol. 28, Issue 1, pp.125-140. Springer.

Coleman, J.S. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.

Colman, A. M. (1982). *Game Theory and Experimental Games*, Oxford: Pergamon.

Colman, A. M. (1999). *Game Theory and its Applications in the Social and Biological Sciences*, 2nd ed., Routledge, New York.

Colman, A.M. (2005). "Game Theory", *Encyclopedia of Statistics in Behavioral Science*, Vol. 2, pp. 688–694, (Eds.) Everitt, B.S. & Howell, D.C., John Wiley, Chichester.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* **(***Essay on the Application of Analysis to the Probability of Majority Decisions***)**. Paris: Imprimerie Royale.

Cooper, R.W., DeJong, D.V., Forsythe, R., & Ross, T.W. (1990). "Selection Criteria in Coordination Games: Some Experimental Results". In: *The American Economic Review*, Vol. 80, No. 1, pp. 218- 233.

Cooper, R., De Jong, D. V., Forsythe, R., & Ross, T. W. (1992). "Forward Induction in Coordination Games". In: *Economics Letters*, *40*(2), 167-172. https://doi.org/10.1016/0165-1765(92)90217-M

"Cooperation". (2020). *Lexico.com (Oxford)*. https://www.lexico.com/definition/cooperation. Accessed on 25.05.2020.

Copi, I.M. (1996). *Symbolic Logic*, 5th Ed., Prentice-Hall, New Delhi.

Copi, I.M. & Cohen, C. (1997). *Introduction to Logic*, 9th ed., Prentice-Hall, New Delhi.

Copi, I.M. & Cohen, C. (2004). *Introduction to Logic*, 11th ed., Prentice-Hall, New Delhi.

Copi, I.M., Cohen, C., & McMahon, K. (2014). *Introduction to Logic*, 14th ed., Harlow: Pearson.

Cornman, J. W., & Lehrer, K. (1974). *Philosophical Problems and Arguments*, 2nd ed. Macmillan, New York.

Crowley, P., 1996, Evolving Cooperation: Strategies as Hierarchies of Rules. *BioSystems*, 37, 67-80.

Crowley, P., and others, 1996, Evolving Cooperation: The Role of Individual Recognition. *BioSystems*, 37, 49-66.

Darwin, C. R. (1859). *On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* 1st Ed., London: John Murray.

Darwin, C. R. (1866). *On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life,* 4th Ed., London: John Murray.

Darwin, C.R. (1871). *The Descent of Man and Selection in Relation to Sex*. London: John Murray.

Davis, J. M. (1958). "The Transitivity of Preferences", In *Behavioral Science*, Vol. 3, pp. 26-33.

Dawes, R. M. (1980). "Social Dilemmas". *Annual Review of Psychology*, 31, 169-193.

Dawes, R.M. & Thaler, R.H. (1988). "Anomalies: Cooperation", *Journal of Economic Perspectives*, Vol. 2, No. 3, pp. 187–197.

Dawkins, R. (2006/2010 reprint). *The Selfish Gene*, 30th Anniversary Ed. Oxford University Press, New Delhi, India.

Deutsch, M.(2005). "Cooperation and Conflict*", The Essentials of Teamworking: International Perspectives*, (Eds.) West, M.A., Tjosvold, D., and Smith, K.G. John Wiley & Sons, Ltd., West Sussex, England.

Dissanayake, S. & Bezwada, N. (2010). *Characteristics and Contributory Causes Related to Large Truck Crashes (Phase I) -Fatal Crashes*, A Report on Research Sponsored by Mid-America Transportation Center, University of Nebraska-Lincoln, USA.

Dixit, A.K. & Nalebuff, B.J. (1991). *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*, Norton & Co., New York.

Dong, L., Montero, M., & Possajennikov, A. (2018). "Communication, Leadership and Coordination Failure". In: *Theory Decision*. 84:557–584. https://doi.org/10.1007/s11238-017-9617-9

Edwards, W., Miles, R. F., & von Winterfeldt, D. (2007). "Introduction". In W. Edwards, R. F. Miles, & D. von Winterfeldt (Eds), *Advances in Decision Analysis: From Foundations to Applications*. New York: Cambridge University Press. pp. 1-12.

Ellers, J., & Pool, N.C.E. van der. (2010). "Altruistic behavior and cooperation: the role of intrinsic expectation when reputational information is incomplete", *Evolutionary Psychology*, 8(1):37-48.

Elster, J. & Hylland, A. (1986). J. "Introduction", *Foundations of Social Choice Theory*. (Eds.) J. Elster & A. Hylland, Cambridge University Press, New York

Elster, J. (1985). *Making Sense of Marx*. New York: Cambridge University Press. (**Def of Methodlgcl Indvdlsm**: P. 5)

Elster, J. (1989/1999). *Nuts and Bolts for the Social Sciences*. Cambridge University Press, Cambridge.

English Oxford Dictionaries. Accessed on 19.11.2017. https://en.oxforddictionaries.com/definition/synergy.

Evans, J. (2014). "Rationality and the Illusion of Choice". In: *Front. Psychol*. 5: 104. Published online 12 Feb 2014. doi: 10.3389/fpsyg.2014.00104.

Feder, H. M. (1996). "Cleaning symbioses in the marine environment". In S. M. Henry (ed.), *Symbiosis*, Vol. 1, pp. 327–380. Academic Press, N.Y.

Fehr, E., Fischbacher, U., Gächter, S. (2002). "Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms". In: *Human Nature*, Vol. 13, No. 1, pp. 1-25.

Fehr, E. & Fischbacher, U. (2003). "The Nature of Human Altruism"; in NATURE, Vol. 425, 23 Oct. 2003, pp. 785-791.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Forber, P. & Smead, R. (2015). "Evolution and the Classification of Social Behavior"; in *Biology and Philosophy*, 30(3):405–421, Springer Netherlands.

French, S. (1986) Decision Theory: An Introduction to the Mathematics of Rationality. Ellis Horwood, Chichester.

Fudenberg, D. & Maskin, E. (1990). Evolution and cooperation in noisy repeated games. *Am. Econ. Rev*., 80: 274–279.

Fudenberg, D. & Tirole, J. (1991). *Game Theory*, MIT Press, Cambridge, MA.

Gallo, P.S. & McClintock, C.G. (1972). "Cooperative Behavior in Mixed-Motive Games", *Cooperation and Competition: Readings on Mixed Motive Games*, (Eds.) Wrightsman, L.S., O'Connor, J., & Baker, N.J. Wadsworth, Belmont, California. (pp. 8-18).

Gauthier, David, 1986, *Morals by Agreement*, New York: Oxford University Press.

Gauthier, David. "Coordination." *Dialogue* 14 (1975): 195-221.

Gauthier, David. "Rational Cooperation." *Nous* 8 (1974): 53-65.

Gayon, J. (Transl. M. Cobb). (1998). *Darwinism's Struggle for Survival: Heredity and the Hypothesis of Natural Selection*. Cambridge Univ. Press, Cambridge, U.K., p. 72.

Gibbard, Allan, 1990, *Utilitarianism and Coordination*, Garland Publishing, London.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). "Explaining altruistic behavior in humans", *Evolution and Human Behavior*, Vol. 24, Pp. 153–172.

Green, S. L. (2002). *Rational Choice Theory: An Overview*. Waco, TX: Baylor University. Retrieved from http://business.baylor.edu/steve_green/green1.doc.

Grether, D.M. & Plott, C. R. (1979): "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, 69, 623-638.

Grim, Patrick, Gary Mar, and Paul St. Denis, *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*, © MIT Press/ Bradford Books 1997.

Guttag, J.V., Horning, J.J., Garland, S.J., Jones, K.D., Modet, A., & Wing, J.M. (1993). A Little Bit of Logic. In: *Larch: Languages and Tools for Formal Specification*. Springer, New York, NY.

Hahn, F. & Hollis, M. (Eds.). (1979). *Philosophy and Economic Theory*, Oxford: OUP.

Haji, Ishtiyaque. "The Compliance Problem." *Pacific Philosophical Quarterly* 70 (1989): 105-21.

Haldane, J.B.S. (1932). *The Causes of Evolution*. London: Longmans, Green & Co.

Haldane, J.B.S. (1955). Population genetics. *New Biology*, 18: 34-51.

Hamilton, W.D. (1963). "The Evolution of Altruistic Behavior", *The American Naturalist*, Vol. 97, No. 896, pp. 354-356, The University of Chicago Press, Chicago.

Hamilton, W.D. (1964). "The Genetical Evolution of Social Behaviour. I & II". *Journal of Theoretical Biology*. 7, 1-16, & 17-52.

Hamlin, A. (1986). Ethics, *Economics and the State*. New York: St. Martin's.

Hammack, R. (2018). *Book of Proof*, 3rd Ed., Department of Mathematics & Applied Mathematics, Virginia Commonwealth University Richmond, Virginia.

Hammond, P. J. (1997). “Rationality in economics”, *Rivista Internazionale di Scienze Sociali*, **Anno CV**, 247–288.

Hampton, Jean. “Can We Agree on Morals?” *Canadian J. of Philosophy* 18 (1988): 331-56.

Hampton, S. J. (2009). *Essential Evolutionary Psychology*. Sage Publications Ltd., London. (Ref. to Ch. 1, “Darwin’s Argument and Three Problems: Heritability, Sexual Selection and Altruism”)

Hansson, S.O. & Grüne-Yanoff, T. (2018). "Preferences", In *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/preferences/>.

Hardin, G. (1968). "The Tragedy of the Commons." *Science*. Vol. 162, pp. 1243-1248.

Hardin, R. (1982). *Collective Action*, Baltimore, MD.

Hargreaves-Heap, S.P. & Varoufakis, Y. (1995) *Game Theory: A Critical Introduction*. London: Routledge.

Hargreaves-Heap, S.P. & Varoufakis, Y. (2004). *Game Theory: A Critical Introduction*., 2nd. Ed., London: Routledge.

Hargreaves Heap, S. (1989). *Rationality in Economics*. Oxford: Blackwell.

Harman, Gilbert. “Rationality in Agreement: A Commentary on Gauthier’s *Morals by Agreement*.” *Social Philosophy and Policy* 5 (1988): 1-16.

Harsanyi, J. C. (1982/1990). “Morality and the Theory of Rational Behaviour”. In: *Utilitarianism and Beyond*, Eds. Sen, A. and B. Williams, Cambridge University Press, New York.

Hatcher, D.L. et al. (1990). *Reasoning and Writing: An Introduction to Critical Thinking*, Baker University, Baldwin City, KS.

Hechter, M. & Kanazawa, S. (1997). "Sociological Rational Choice Theory", *Annual Review of Sociology* 23: 191–214.

Hempel, C.G. (1960). Inductive Inconsistencies. *Synthese* **12,** 439–469. https://doi.org/10.1007/BF00485428

Heylighen, F. (1992). "Evolution, Selfishness and Cooperation", Journal of Ideas, Vol. 2, # 4, pp 70-76.

Hitchcock, D. (2011)." Instrumental Rationality". In P. McBurney, I. Rahwan & S. Parsons (Eds.), *Argumentation in Multi-Agent Systems*. Proceedings, 7th International Workshop, ArgMAS 2010 (pp. 1-11). New York/Heidelberg: Springer.

Hobbes, T. (1651). *Leviathan*. Printed for Andrew Crooke, at the Green Dragon in St. Paul's Church-yard, London.

Hobbes, T. (1651, 1991). *Leviathan*. (Ed.) R. Tuck. Cambridge: Cambridge University Press.

Hobbes, T. (1651/1985). *Leviathan*. (Ed.) C. MacPherson. London: Penguin.

Hooker, J. N. (2013). "Moral Implications of Rational Choice Theories". In C. Lütge (Ed.), *Handbook of the Philosophical Foundations of Business Ethics* (pp. 1459-1476). Springer.

Hume, D. (1740/1888). *Treatise of Human Nature*, (Ed.) L.A. Selby-Bigge. Oxford: Oxford University Press.

Hurley, P.J. (2012). *A Concise Introduction to Logic*, 11th Ed., Wadsworth, Boston, MA.

Imhof, L.A., Fudenberg, D., & Nowak, M.A. (2007). "Tit-for-tat or Win-stay, Lose-shift?", *J. Theor. Biol*. 247(3): 574–580.

Irwin, T. (2009). *Implications for Climate-Change Policy of Research on Cooperation in Social Dilemmas*. Policy Research Working Paper 5006, Background Paper to the 2010 World Development Report, The World Bank.

Jeffrey, R. (1981). Formal Logic: Its Scope and Limits (2nd ed.). New York: McGraw-Hill.

Johansson, P.O. & Löfgren, K.G. (2003). "Economic Efficiency". In: A. Kuper & J. Kuper (Eds.), *The Social Science Encyclopedia*, 2nd Ed., p.217, New York: Routledge.

Julian Nida-Rümelin. (1997). *Economic Rationality and Practical Reason*, Kluwer Academic Publishers,

Jurišić, M., Kermek, D., & Konecki, M. (2012). "A review of iterated prisoner's dilemma strategies", *Proceedings of the 35th international conventionon in formation and communication technology, Electronics and Microelectronics* (MIPRO, May 21-25, 2012, Opatija, Croatia). pp. 1093–1097.

Kalberg, S. (1980). "Max Weber's Types of Rationality: Cornerstones for the Analysis of Rationalization Processes in History", In: *The American Journal of Sociology*, Vol. 85, No. 5, pp. 1145-1179.

Kahneman, D. & Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk". *Econometrica*, 47, 263–291.

Kahneman, D. (2003). "Maps of Bounded Rationality: Psychology for Behavioral Economics". In: *The American Economic Review*, Vol. 93, No. 5, pp. 1449-1475. <https://www.jstor.org/stable/3132137>

Kahneman, D. (2011). Thinking, Fast and Slow. New Delhi: Penguin Books.

Kapeller, J., Schütz, B., & Steinerberger, S. (2013). "The Impossibility of Rational Consumer Choice: A Problem and Its Solution", in ***J. Evol. Econ.*** Vol. 23, pp. 39–60. DOI 10.1007/s00191-012-0268-2.

Kaufman, S.B., & Jauk, E. (2020). "Healthy Selfishness and Pathological Altruism: Measuring Two Paradoxical Forms of Selfishness". In: *Frontiers in Psychology*, Front. Psychol. 11:1006. doi: 10.3389/fpsyg.2020.01006

Kavka, Gregory. "Morals by Agreement." *Mind* 96 (1987): 117-21.

Keddy, P.A. 2001. *Competition*, 2nd ed., Kluwer, Dordrecht. 552 p.

Keef, P. & Guichard, D. (2020). *An Introduction to Higher Mathematics*, Creative Commons, https://www.whitman.edu/mathematics/higher_math_online/higher_math.pdf

Kelley, D. (2014). *The Art of Reasoning: An Introduction to Logic and Critical Thinking*, 4th Ed., New York: Norton.

Kelly, A. (2003). *Decision Making Using Game Theory: An Introduction for Managers*, Cambridge University Press, New York.

Kelly, T. (2003). "Epistemic Rationality as Instrumental Rationality: A Critique". In: *Philosophy and Phenomenological Research*, Vol. 66, No.3, pp. 612-640.

Kin selection. (n.d.). In *Wikipedia, The Free Encyclopedia*. Retrieved January 30, 2018, from <https://en.wikipedia.org/wiki/Kin_selection>

King, B. (2015). "The Prisoner's Dilemma and the Evolution of Morality", *Philosophy Now*, Vol. 109, Aug./Sep 2015, Pages 14-17. <https://philosophynow.org/issues/109/The_Prisoners_Dilemma_and_The_Evolution_of_Morality>

Kitcher, P., 1993, The Evolution of Human Altruism, *Journal of Philosophy*, 90, 497-516.

Klarreich, E. (2017, July 18). “In Game Theory, No Clear Path to Equilibrium”. In: *Quanta Magazine*. https://www.quantamagazine.org/in-game-theory-no-clear-path-to-equilibrium-20170718/

Kohn, Alfie (1992). *No Contest: The Case Against Competition*. Houghton Mifflin Co., N.Y., p. 19.

Kollock, P. (1998). “Social Dilemmas: The Anatomy of Cooperation”. In: *Annual Review of Sociology*, Vol. 24, pp. 183-214.

Krockow, E. (2020). *Why Do People Act Against Their Own Better Judgement?*. Retrieved on 10/06/2020 from <https://www.psychologytoday.com/us/blog/stretching-theory/202006/why-do-people-act-against-their-own-better-judgement>

Social Dilemmas: The Anatomy of Cooperation Author(s): Peter Kollock Source: Annual Review of Sociology, Vol. 24 (1998), pp. 183-214

Kothari, C.R. (2012). *Research Methodology: Methods and Techniques*, 2nd Ed., New Age International Publishers, New Delhi.

Kreps, D.M. (1990). *A Course in Microeconomic Theory*. Princeton, N.J. : Princeton University Press.

Kuhn, S. (2017). "Prisoner's Dilemma", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/prisoner-dilemma/>.

Lea, S.E.G., Tarpy, R.M., & Webley, P. (1987). *The Individual in the Economy: A Textbook of Economic Psychology*. New York: Cambridge University Press.

Leite, A. (2007). "Epistemic Instrumentalism and Reasons for Belief: A Reply to Thomas Kelly's 'Epistemic Rationality as Instrumental Rationality: A Critique'", *Philosophy and Phenomenological Research*, Vol. 75, No. 2, pp. 456-464.

Li, S. (2006). "Preference Reversal: A New Look At an Old Problem". In: *The Psychological Record*, Vol. 56, No. 3, pp. 411-428.

Lichtenstein, S. & Slovic, P. (1971): "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, 89, 46-55.

Liebrand, W. (1983). "A Classification of Social Dilemma Games". In: *Simulation & Games*, Vol. 14, pp. 123–138.

Lipsey, R.G. (1968). "Introduction", in: *An Introduction to Positive Economics*, 2nd Ed., London: ELBS.

Locke, J. (1689). **Second Treatise of Government.**

Longman Dictionary of Contemporary English, Longman Group Ltd., 3rd Ed., 1995 (2nd Indian reprint, 2000), p.224

Lorimer, P., 1967. "A note on orderings". *Econometrica* 35:537-539.

Luce, R. D., Raiffa, H., 1957, *Games and Decisions*, Wiley, New York.

Machlup, F. (1965). "Why Economists Disagree." In: *Proceedings of the American Philosophical Society*, Vol. 109, pp. 1–7.

Machol, R.E. (1965). *System Engineering Handbook*. McGraw-Hill, New York.

Macionis, J.J. (2000). *Society: The Basics*, 5th Ed., Prentice Hall, New Jersey.

Mankiwi, N.G. (2018). *Principles of Economics*, 8th ed., Cengage, Boston, MA.

McWilliams, N. (1984). "The Psychology of the Altruist". In: *Psychoanalytic Psychology*, Vol. 1, No. 3, pp. 193-213.

Malthus, T. R. (1798). *An Essay on the Principle of Population, as It Affects the Future Improvement of Society*. J. Johnson, St. Paul's Church-Yard, London.

Margolis, H. (1982). *Selfishness, Altruism, and Rationality*. Cambridge: Cambridge University Press.

Mas-Collel, A., Whinston, M.D., & Green, J.R. (1995). *Microeconomic Theory*. New York: Oxford University Press.

Maynard Smith, J. (1982). Evolution and the Theory of Games, Cambridge University Press, Cambridge.

McElreath, R. & Boyd, R. (2007). *Mathematical Models of Social Evolution*: *A Guide for the Perplexed*. Univ. of Chicago Press, Chicago. (p. 82)

McFarland, D. & Bösser, T. (1993). *Intelligent Behavior in Animals and Robots*. MIT Press, London, England. (Ch. 2, Rational Behavior, Pp. 25-39)

Mele, A.R. & Rawling, P. (2004). "Introduction: Aspects of Rationality". In A. R. Mele, & P. Rawling (Eds.), *The Oxford Handbook of Rationality* (1st ed., pp. 3-13). New York: Oxford University Press.

Melis, A.P. & Semmann, D. (2010). "How is human cooperation different?", in *Phil. Trans. R. Soc. B*, Vol. 365, pp. 2663–2674. doi:10.1098/rstb.2010.0157.

Mendel, J.G. (1865). *Experiments in Plant Hybridization*.

Mikkelsen, P.M. & Robin, H. (2011). *The Teacher-Friendly Guide to Evolution*. PRI Special Publication no. 40, NY, USA, URL=<http://teacherfriendlyguide.org/bivalves>

Mill, J.S. (1882). *A System of Logic*, 8th Ed., New York: Harper & Brothers.

Mises, L. von. (1996). *Human Action: A Treatise on Economics*. 4th ed., Fox & Wilkes, San Francisco, CA. (Mises 1996: 42-43) (Appraisal of MethdInd vs MethdCollect, pp. 42-43.)

Molander, P., 1985, The Optimal Level of Generosity in a Selfish, Uncertain, Environment, *Journal of Conflict Resolution*. 29, 611-618.

Moon, T., Frost, R., and Stirling, W., 1996, An Epistemic Utility Approach to Coordination in the Prisoner's Dilemma, *BioSystems*, 37, 167-176.

Morgan, C.T., King, R.A., Weisz, J.R., & Schopler, J. (1986). *Introduction to Psychology*, 7th ed. McGraw-Hill Book Co., Singapore.

Myerson, R.B. (2013). *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, MA.

Mutalik, P. (2017, Sept. 14). "Are Genes Selfish or Cooperative?", *Quanta Magazine*, https://www.quantamagazine.org/are-genes-selfish-or-cooperative-20170914/

Nagel, T. (1970). *The Possibility of Altruism.* Princeton: Princeton University Press.

Nash, J. (1950). "Equilibrium Points in n-person Games". In: *Proceedings of the National Academy of Sciences*, USA, **36**: 48–49.

Nash, J. (1951). "Non-cooperative Games". In: *Annals of Mathematics*, **54**: 286–295.

Neusner, J. & Chilton, B.D. (2005). *Altruism in World Religions*. (Eds.) Neusner, J. & Chilton, B.D., Washington DC: Georgetown University Press.

Nisbet, R.A. (2008). "Cooperation", *International Encyclopedia of the Social Sciences*. . Retrieved March 25, 2017 from Encyclopedia.com: http://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/cooperation

Noë, R. (2017). "Cooperation", in Stein, J. (ed.). *Encyclopedia of Behavioral Neuroscience*, Vol. **1**, pp. 1-9. Elsevier Inc., Amsterdam.

Nowak, M., May, R. (1992). "Evolutionary Games and Spatial Chaos". In: *Nature*, **359,** 826-829. https://doi.org/10.1038/359826a0

Nowak, M. & K. Sigmund. (1992). Tit for tat in heterogeneous populations. *Nature,* 355:250-253.

Nowak, M. & K. Sigmund. (1993). A strategy of win-shift, lose-stay that outperforms tit-for-tat in the Prisoners' Dilemma game. *Nature,* 364:56-57.

Nowak, M.A. (2006). "Five rules for the evolution of cooperation". *Science*, 314, 1560–1563.

Nowak, M.A. (2012). "Evolving cooperation", in *J. of Theo. Biology*, Vol. 299, pp. 1–8.

Oakley, B., Knafo, A., & McGrath, M. (2011). "Pathological Altruism—An Introduction". In: B. Oakley, A. Knafo, G. Madhavan, & D. S. Wilson (Eds.), *Pathological Altruism*. (pp. 3-9). New York, NY: Oxford University Press.

Oakley, B. A. (2013). "Concepts and Implications of Altruism Bias and Pathological Altruism". In: *Proc. Natl. Acad. Sci. U.S.A.* Vol. 110(Suppl. 2), pp. 10408–10415. In: PNAS, pp. 1-8, downloaded on 16 May 2020 from www.pnas.org/cgi/doi/10.1073/pnas.1302547110

Okasha, S. (2006). *Evolution and the Levels of Selection*. Oxford University Press. 10.1093/acprof:oso/9780199267972.001.0001

Okasha, S. (2013). "Biological Altruism", In: *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2013/entries/altruism-biological/>.

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard U. Press.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press, 1990.

Oxford English Dictionary (2001). (Eds.) Simpson, J. & Weiner, E., 2nd ed., Vol. XIII,. Clarendon, Oxford, (1989/2001), [For Def. of "Rationality", 2nd ed., Vol. XIII, p. 220.]

Parsons, T. (1937). *The Structure of Social Action*, McGraw Hill, New York, p. 58.

Pavitt, C. (2018). "The Path to Cooperative Action during Group Social Dilemmas: A Literature Review, Set of Propositions, and Model Describing How the Opportunity to Communicate Encourages Cooperation". In: *The Review of Communication Research*, 6, 54-83. https://doi.org/10.12840/issn.2255-4165.2018.06.01.016.

Perloff, J.M. (2000). *Microeconomics*, 6th ed., Addison-Wesley, Boston, MA

Płatkowski,T. (2017). "On Derivation and Evolutionary Classification of Social Dilemma Games". In: *Dyn Games Appl*, Vol. 7, pp. 67–75. DOI 10.1007/s13235-015-0174-y.

Plato. (1941). *The Republic*. (Trans.) F.M. Cornford, London: Oxford University Press.

Popper, K.R. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Popper, K.R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books.

Posner, R.A. (1997). "Rational Choice, Behavioral Economics, and the Law". 50 Stanford Law Review 1551.

Prisner, E. (2014). *Game Theory Through Examples*. The Mathematical Association of America, Washington, DC.

"Proof." *Merriam-Webster.com Dictionary*. Merriam-Webster, https://www. merriam-webster.com/dictionary/proof. Accessed 13 May. 2020.

Provis, C. "Gauthier on Coordination." *Dialogue* (Canada) 16 (1977): 507-9.

Pulford, B.D., Colman, A.M., & Lawrence, C.L. (2014). "Strong Stackelberg Reasoning in Symmetric Games: An Experimental Replication and Extension". *PeerJ* 2: e263 https://doi.org/10.7717/peerj.263

Putnam, H. (2002). The Collapse of the Fact/Value Dichotomy. Cambridge, MA: Harvard University Press.

Rachels, J. (2003). *The Elements of Moral Philosophy*, 4th ed. McGraw-Hill, New York, NY.

Raiffa, H. (1968). *Decision Analysis*, Reading, Mass.:Addison-Wesley, , pp. 78-79.

Rand, D.G., & Nowak, M.A. (2013). *Trends in Cognitive Sciences*, Vol. 17, No. 8, pp. 413-425.

Rapoport, A. & Chammah, A. M. (1965). *The Prisoner's Dilemma*. Ann Arbor, MI: University of Michigan Press.

Rapoport, A. & Chammah, A. M. (1966). *The Game of Chicken*. American Behavioral Scientist, 10(3), 10-28. DOI: 10.1177/000276426601000303

Rapoport, A. (1987). "Prisoner's Dilemma", *The New Palgrave: A Dictionary of Economics*, 1st Ed., (eds.) Eatwell, J., Milgate, M., & Newman, P. New York, Norton.

Rapoport, A. (1989). "Prisoner's Dilemma", *The New Palgrave: Game Theory*, (eds.) Eatwell, J., Milgate, M., & Newman, P. New York, W. W. Norton.

Rapoport, A., Seale, D.A., & Colman, A.M. (2015). IsTit-for-Tat the Answer? On the Conclusions Drawn from Axelrod'sTournaments. *PLOS ONE* 10(7): e0134128. doi:10.1371/journal.pone.0134128.

Raub, W., Buskens, V., & Corten, R. (2015). "Social Dilemmas and Cooperation". In: Braun, N., Saam, N.J. (eds.) Handbuch Modellbildung und Simulation in den Sozialwissenschaften. Springer VS, Wiesbaden. DOI 10.1007/978-3-658-01164-2_21.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Reciprocal altruism. (n.d.). In *Wikipedia, The Free Encyclopedia*. Retrieved January 30, 2018, from <https://en.wikipedia.org/wiki/Reciprocal_altruism>

Regan, Donald, 1980, *Utilitarianism and Cooperation*, Oxford.

Regenwetter, M., Dana, J., & Davis-Stober, C.P. (2011). "Transitivity of Preferences". In *Psychological Review*, Vol. 118, No. 1, pp. 42–56. American Psychological Association, DOI: 10.1037/a0021150

Resnik, M. (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press, Minneapolis, MN.

Ridley, M. (1996). *The origins of virtue*. London: Viking.

Ridley, M. (2004). *Evolution*, 3rd ed. Blackwell, Malden, USA.

Riegelman, R. (1979). "Contributory Cause: Unnecessary and Insufficient". In: "*Postgrad Med*". Vol. 66, No. 2, 177-9.

Riegelman, R.K. (2012). *Studying a Study and Testing a Test: Reading Evidence-based Health Research*, 6th Ed., Wolters Kluwer Health, New York.

Riker, W.H. (1982). *Liberalism against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. San Francisco: W.H. Freeman. (p. 2).

Roberts, G. & Renwick, J.S. (2003). The development of cooperative relationships: an experiment, *Proc. R. Soc. B.* 270(1530):2279-83.

Rousseau, J.J. (1762). *The Social Contract*.

Rousseau, J. (1984). *A Discourse on Inequality*. Trans. M. Cranston. New York: Penguin Books.

Russell, B. (1903). *The Principles of Mathematics*, Cambridge University Press, London.

Russell, Bertrand. 1951. *New Hopes for a Changing World*. George Allen & Unwin Ltd., London.

Russell, B. (1954/1963). *Human Society in Ethics and Politics*, London: George Allen & Unwin, (3rd impression 1963).

Russell, B.W. (1959). *Common Sense and Nuclear Warfare*. George Allen and Unwin, London.

Russell, B. (2009). "An Outline of Intellectual Rubbish", *The Basic Writings of Bertrand Russell*. (eds.) Egner, R.E. & Denonn, L.E., Routledge, London.

Ryle, G. (1949). *The Concept of Mind*. Hutchinson, London.

Salgado, M., Noguera, J.A., & Miguel, F.J. (2015). Modelling Cooperation Mechanisms: Some Conceptual Issues, *Journal of Archaeological Method and Theory*, 21(2):325-342.

Samuels, R., Stich, S., & Faucher, L. (2004). "Reason and Rationality", *Handbook of Epistemology*. (Eds. Niiniluoto, I., Sintonen, M., & Wolenski, J.). Dordrecht: Kluwer. (Pp. 1-50).

Samuelson, P.A. & Nordhaus, W.D. (2006). *Economics*, 18th ed., Tata McGraw-Hill, India.

Savage, L. J. (1954). The Foundations of Statistics. (2nd ed.). New York: John Wiley.

Schloss J.P. (2017). "Darwinian Explanations of Morality: Accounting for the Normal but not the Normative", *Understanding Moral Sentiments: Darwinian Perspectives?* (Eds.) Putnam, H., Neiman, S., & Schloss, J.P. Routledge, London, 2017, pp.81-121. <http://isthmussociety.org/Documents/schloss_reading.pdf> Accessed on 16.8.2018.

Seelig, B.J. & Rosof, L. (2001). "Normal and Pathological Altruism", *Journal of the American Psychoanalytic Association*, Vol. 49, No. 3, pp. 933 – 959.

Sen, A. K. (1977a), *'Rational Fools'*, *Philosophy and Public Affairs*, 6, 317-44.

Sen, A. K. (1977b). 'Rationality and Morality: A Reply', *Erkenntnis* 11.

Sen, A.K. (1970). *Collective Choice and Social Welfare*. San Francisco, CA: Holden-Day.

Sen, Amartya K., 1974, 'Choice, Orderings and Morality', *Practical Reason*, ed., S. Korner, Oxford: Blackwell.

Sen, A.K. (1987/reprint 2004). *On Ethics and Economics*. Oxford: Blackwell.

Sen, A.K. (1990). "Rational Behavior". In: Eatwell, J., Milgate, M., & Peter, N., *Utility and Probability*, (New York: W. W. Norton & Company), pp. 1998-216.

Shankar, A. & Pavitt, C. (2002). "Resource and Public Goods Dilemmas: A New Issue for Communication Research". In: *The Review of Communication*, 2 (3), 251 – 272.

Simon, H.A. (1972). "Theories of Bounded Rationality". In: McGuire, C.B. & Radner, R. (Eds.), *Decision and Organization*, (pp. 161-176). Elsevier, Amsterdam.

Simon, H. A. (1976) "From Substantive to Procedural Rationality", in Spiro J. Latsis, *Method and Appraisal in Economics*, Cambridge: Cambridge University Press: 129-148 (The Quote is from p. 132.

Simon, H.A. (1987). "Bounded Rationality". In: John Eatwell, Murray Milgate, & Peter Newman (Eds.), *The New Palgrave*: *A Dictionary of Economics*, Vol. I (pp. 266-268). New York: Palgrave.

Singer, Emily. (February 12, 2015). "Game Theory Calls Cooperation into Question", *QUANTA MAGAZINE*. (https://www.quantamagazine.org/20150212-game-theory-calls-cooperation-into-question/).

Singer, P. (2011). *Practical Ethics*, 3rd ed., Cambridge University Press, New York, NY, USA.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, New York.

Smith, A. (1776). *The Wealth of Nations.*

Smith, A. (1759). *The Theory of Moral Sentiments*. London, UK: A. Millar.

Sobel, J. H. "Interaction Problems for Utility Maximizers." *Canadian J. of Phil.* 4 (1975): 677-88

Sober, E. & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard UP, Cambridge, MA.

Stein, E. (1997). "Can We Be Justified in Believing That Humans Are Irrational?". In: *Philosophy and Phenomenological Research*, Vol. 57, No. 3, pp. 545-565.

Stich, S. (1999). "Is Man a Rational Animal?" In *Questioning Matters: An Introduction to Philosophical Inquiry*, ed. D. Kolak. Mountain View, Calif.: Mayfield Publishing, pp. 221-236.

Straffin, P. (1980). "The Prisoner's Dilemma". In: *UMAP Journal*, **1**, 101-103.

Straffin, P.D. (1993). *Game Theory and Strategy*, The Mathematical Association of America, Washington, D.C.

Su, Q., Li, A., Wang, L., & Stanley, H.E. (2019). "Spatial Reciprocity in the Evolution of Cooperation". In: *Proc. R. Soc. B*. 286: 20190041. http://dx.doi.org/10.1098/rspb.2019.0041

Sun, S. (2018). "From Defensive Altruism to Pathological Altruism". In: *SAGE Open*. pp. 1–8. DOI: 10.1177/2158244018782585 journals.sagepub.com/home/sgo

Suppes, P. (1967). 'Decision Theory', in P. Edwards (ed.), *The Encyclopedia of Philosophy*, N. Y.: Macmillan.

Sussman, R.W. & Cloninger C.R. (2011). "Introduction: Cooperation and Altruism". In: Sussman, R. & Cloninger, C. (eds), Origins of Altruism and Cooperation. Series: Developments in Primatology: Progress and Prospects, Vol. 36. Springer, New York, NY (Quote is from Page 2).

Svavarsdottir, S. (2008). "The Virtue of Practical Rationality". In: *Philosophy and Phenomenological Research*, Vol. 77 (1): 1-33. /(Vol. 77, No. 1, pp. 1-33. (Quote from p. 1)

Swartz, N. (1997). *The Concepts of Necessary Conditions and Sufficient Conditions*. <http://www.sfu.ca/~swartz/conditions1.htm>

Symmetric game. (n.d.). In *Wikipedia, The Free Encyclopedia*. Retrieved May 30, 2018, from <https://en.wikipedia.org/wiki/Symmetric_game>

Taylor, M. (1976). *Anarchy and Cooperation*. London: John Wiley and Sons.

Taylor, P.W. (1974). *Principles of Ethics: An Introduction*. Wadsworth, Belmont, California.

Tilley, J.J. (1991). "Altruism and the Prisoner's Dilemma". In: *Australasian Journal of Philosophy*. Vol. **69**, No. 3, pp. 264–287.

Tomasello, M. (2009). *Why We Cooperate*. MIT Press, Cambridge, MA.

Trivers, R.L. (1971). "The evolution of reciprocal altruism". *Quarterly Review of Biology*. 46: 35–57. doi:10.1086/406755

Trivers, R.L. (1985). *Social Evolution*. Benjamin/ Cummings Pub. Co., Menlo Park.

Tucker, A.W. (1950). "A two-person dilemma" (Mimeographed paper, Stanford University), in Rasmusen, E.B., ed., *Readings in Games and Information* (1989), 7-8. Blackwell Publishers: Oxford.

Tuomela, R. (2000). *Cooperation: A Philosophical Study*, Kluwer Academic Publishers, Dordrecht, the Netherlands.

Tversky, A. (1969). "Intransitivity of Preferences". In *Psychological Review,* Vol. 76, No. 1, pp. 31–48. https://doi.org/10.1037/h0026750

Tversky , A. & Thaler, R.H. (1990). "Anomalies: Preference Reversals", *Journal of Economic Perspectives*, Vol. 4, Pp. 201–21.

Tversky, A. & Kahneman, D. (1992). "Advances in Prospect Theory: Cumulative Representation of Uncertainty". *Journal of Risk and Uncertainty*, 5, 297–323.

Tversky, A., & Kahneman, D. (2002). Judgment under Uncertainty: Heuristics and Biases. In: D.J. Levitin (Ed.), *Foundations of Cognitive Psychology: Core Readings* (p. 585–600). MIT Press.

Tyran, J.R. (2020). "Coordination Failure". In: *Encyclopedia.com*. Retrieved on May 28, 2020 from https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/coordination-failure

Understanding Evolution. (2016). Univ. of California Museum of Paleontology. Accessed on 25 August 2016, URL= <http://evolution.berkeley.edu/>.

Vallentyne, Peter, (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*, Cambridge U. Press, New York, 1991.

van Damme, E. (1995). *Game Theory: The Next Stage*, Working Paper, Revised Version, March, 1995, Download date: 19. Sep. 2015.

Van Huyck, J.B., Battalio, R.C., & Beil, R.O. (1990). "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure". In: *American Economic Review,* 80, 234–248.

Varian, H.R. (1992). *Microeconomic Analysis*, 3rd ed., New York, N.Y.

Varian, H.R. (2010). *Intermediate Microeconomics: A Modern Approach*, 8th Ed., Norton, New York, N.Y.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Weber, M. (1922/ 1978). *Economy and Society: An Outline of Interpretive Sociology*. Berkley, CA: U. California Press.

Werner Leinfellner & Eckehart Köhler (Eds.), *Game Theory, Experience, Rationality*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, 472 pp.

Weston, S.C. (1994). "Toward a Better Understanding of the Positive/Normative Distinction in Economics". *Economics and Philosophy*, Vol. 10, pp. 1-17.

Whitehead, A.N. & Russell, B. (1927/ 1963). *Principia Mathematica*, Vol. I, 2nd Ed., Cambridge University Press, London.

Wilson, D.S. (1992). "On the Relationship between Evolutionary and Psychological Definitions of Altruism and Selfishness." In: *Biology and Philosophy* **7**: 61-68. The Netherlands: Kluwer.

Wilson, E.O. (2014). *The Meaning of Human Existence*, 1st ed. Liveright, New York.

Wit, S. & Dickinson, A. (2009). Associative Theories of Goal-Directed Behaviour, *Psychological Research*, Springer, 73, 463–476.

Wolfram, S. (1994). *Philosophical Logic: An Introduction*, London: Routledge.

Wu, J. & Axelrod, R. (1995). How to Cope with Noise in the Iterated Prisoner's Dilemma. *J. Conflict Res.*, Vol. 39, pp. 183–189.

Wynne-Edwards, V. C. (1962). *Animal Dispersion in Relation to Social Behaviour*. Oliver & Boyd, Edinburgh.

Yia, S.D., Baekb, S.K., & Choic, J.K. (2016). "Combination with Anti-Tit-for-Tat Remedies Problems of Tit-for-Tat", *Journal of Theoretical Biology*, 412, Sept. 2016.

Zaggl, M.A. (2014). "Eleven mechanisms for the evolution of cooperation", in *Journal of Institutional Economics*, 10: 2, 197–230.

Zahavi, A. (1975). Mate Selection: A Selection for Handicap. *Journal of Theoretical Biology*, **53**, 205–214.

Zahavi, A. (1995). "Altruism as a Handicap: The Limitations of Kin Selection and Reciprocity", *Journal of Avian Biology*, Vol. 26, No. 1, pp. 1-3. <http://www.jstor.org/stable/3677205> Accessed: 29-01-2018 22:45 UTC.

Zey, M. (1998). *Rational Choice Theory and Organizational Theory: A Critique*. Sage Publications, Thousand Oaks, California.

Kohlberg, L. (1981). The philosophy of moral development. San Francisco: Harper & Row.
Kohlberg, L. (1984). The psychology of moral development. San Francisco: Harper & Row.

Aristotle. (1941). *Nicomachean Ethics*, trans. W.D. Ross. Random House, New York. (Bk.1, Ch.13)

Stich, S. (Accessed on 5 Nov., 2006). Is Man a Rational Animal? <http://www.rci.rutgers.edu/~stich/104_Master_File/104_Readings/Stich/Rational_Animal.pdf>

Arrow, K.J. (1963). *Social Choice and Individual Values*, 2nd Ed. John Wiley, NY.

Bell, D.E., Raiffa, H., & Tversky, A. (1988). "Descriptive, Normative, and Prescriptive Interactions in Decision Making", in *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Bell, D.E., Raiffa, H., & Tversky, A. (Eds.). NY: Cambridge University Press. (Ch. 1, Pp. 9-30)

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1975). Reflections of Language. New York: Pantheon Books.

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, Ma: MIT Press.