UNIVERSITY OF SOUTHAMPTON

"ESSAYS ON NORMS AND GROWTH IN A DYNAMICAL PERSPECTIVE"

BY GIANLUCA FRANCESCO GRIMALDA

DOCTOR OF PHILOSOPHY

FACULTY OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS

FIRST SUBMISSION MARCH 2003

FINAL SUBMISSION SEPTEMBER 2003

UNIVERSITY OF SOUTHAMPTON

"ESSAYS ON NORMS AND GROWTH IN A DYNAMICAL PERSPECTIVE"

BY GIANLUCA FRANCESCO GRIMALDA

DOCTOR OF PHILOSOPHY

FACULTY OF SOCIAL SCIENCES

DEPARTMENT OF ECONOMICS

This thesis was submitted for examination in March 2003. The final version after the required minor amendments were completed is submitted in September 2003.
This thesis is the result of work done mainly while I was in registered postgraduate candidature.

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS

Doctor of Philosophy

ESSAYS ON NORMS AND GROWTH IN A DYNAMICAL PERSPECTIVE

By Gianluca Grimalda

Part 1 develops a model of how social norms come to have motivational power on individual behaviour. This draws on the idea that an individual weighs different 'reasons to action', of which self-interest is only one between them when making decision. These are determined in relation with *internally consistent* standards of assessment of the situation that the agent faces. A formal model is developed, and the existing theories based on the notion of normative expectations are contrasted with a model developed in Chapter 2, in which individuals' other-regarding motivation consists of a conditional willingness to comply with some common ideas of public interest. The two models are contrasted in Chapter 4 within an evolutionary framework, and different concepts of equilibrium and of their stability are put forward in order to appreciate the different implications of the two models in terms of emergence of social norms and of the cognitive and strategic structure enforcing them.

The second part of the thesis develops a model of growth, in which lock-ins to sub-optimal outcomes occur because of 'co-ordination failures' between agents about the choices of the skills by employees and technologies by employers. The model differs from the traditional accounts of growth because of three key assumptions: agents are boundedly rational; prices are not perfectly flexible; a variety of technologies exists, which follow a pattern of technical change of the localized type; this makes technological information a public good only at the sectoral, but not at the aggregate level.

The third part of the thesis develops the earlier analysis and provides a framework in which social norms and economic outcomes can be jointly analysed. A model is developed addressing the question of the relationship between social norms, technology choice, and the optimal institutional management of uncertainty. In particular, two socio-economic settings are available: one requires joint forms of production to agents and uncertainty is managed at the collective level; in the other agents compete between each other, production is individual, thus there is no form of risk-sharing. Various possible scenarios are investigated, and it is argued that the 'economic' and 'social' side reciprocally influence each other. Thus, the relative efficiency of the two types of productive activity affects the type of norms that emerge, and social norms may prevent or facilitate the adoption of productive activities. Hence, social norms can either have the 'progressive' role of a form of 'social capital' for society, but they may also take on a 'conservative' role in preventing the society from switching to better technologies when these become available.

ACKNOWLEDGMENTS

Working on a Ph.D. thesis is such an absorbing experience that, notwithstanding the attempts to confine it to the mere 'professional' side of one's life, inevitably it affects all of the other spheres. I thus feel indebted to any of the people whom I have got to know during my stay in this country that erects cathedrals to science and wisdom. I think I have learnt something from all of them. Even if the list of them is too long to remember here, I will definitely keep a memory of you.

The first person to whom my deepest thanks go is my supervisor Alan Hamlin, whose advices, constructive criticism, and patience represented an invaluable aid throughout my work. Between the many things I have learnt from him, the major one is that, in spite of the overwhelming formalism that is enfolding economics and its applications, the main goal of an economist's research should be to discuss ideas, rather than the scaffoldings of those ideas.

I am also most grateful to the members of the examination committee, Maurice Krueger and Shaun Hargreaves-Heap, for their insightful comments and for the patience shown in going through my 'picturesque', as someone called it, English.

My sincere thanks also go to the economics Department as a whole, starting from the members of the secretarial office, and Gill Crumplin in particular, whose commitment to search a piano for my practicing gave me the ultimate go-ahead to come and study in Southampton.

I also feel indebted to the members of the research and teaching staff, with whom the wine seminars series provided an appropriate setting for perceptive discussions on anything ranging from the reform of the National Health System to the effects of pasteurised milk in cheese preparation on pregnant women. The motto of one of the major director of the event, 'Life is not a picnic', will always be part of my erudition. Between the many, I would just like to mention Antonella Ianni, Xavier

Mateos, Akos Valentinyi and Juuso Valimaki, in whose courses I moved the first steps as a teacher, Giulio Seccia, who acted as a guinea-pig for my piano teaching during a spell, and, last but not least, John Aldrich and Jan Podivinsky, whom I even had the honour to captain during the memorable challenge between Italian wines against rest of the world wines in a special edition of the wine seminar. Long live this commendable tradition!

Furthermore, I wish to express my sincere gratitude to my closest friends, who helped and encouraged me along the way: Colin Jennings and Surjinder Johal, who had a major role in my process of socialization to the English way of life, Alessandra Canepa for being an excellent fellow concert-goer and Grazia Rapisarda for introducing me to the typical 'Comasco' humour.

Finally, I am deeply grateful to my parents and to the person who probably had to suffer the most in terms of the invasion of 'Ph.D.ing' into other spheres: thanks Emma for your 'magic' support and love.

Coventry, 16 September 2003

Gianluca Grimalda

CONTENTS

PREFACE

The topics of social norms and economic growth have both enjoyed renewed interest from economists and other social scientists in recent years. In particular, the revival of 'new' growth theories dates back to the 1980s, i.e. a decade after the dramatic process of steady growth enjoyed in the post-war period by most developed countries had come to an end. This fact, along with the observation of persistent inequalities in per capita level of income across countries and, more strikingly, in their growth rates, delivered fresh questions for economists to answer. This led to the amendment of the classical Solow model in order to get a better grip on reality, which was accomplished through the recognition of an array of 'endogenous' variables, such as human capital, R&D activities, economic policy variables, as capable of accounting for these differences.

Economists' interest in social norms is relatively recent, and can be seen, on the one hand, as the result of a 'natural' process of diffusion of economic techniques of analysis into fields traditionally outside the scope of economics. For instance, in the mid-90s some economists pioneered the use of game theoretical analysis in accounting for social phenomena[1]. On the other hand, though, the growth of economic interest in social norms can also be deemed as an attempt to breach one of the strongholds of modern 'neo-classical' Economics, i.e. rational choice theory, and to replace its most notorious character, i.e. the infamous 'homo-oeconomicus', with an agent more sympathetic to the call of others' interests, alongside her own. Even in this case, adverse 'empirical evidence', in the form of the findings of the newborn *experimental*, or *behavioural, economics*, played a part in calling for a reform of the received theory. Whether this is seen as an 'export' or 'import' of economics knowledge into or from

---

[1] For instance, Kandori (1992), Bernheim (1994), Okuno-Fujiwara and Postlewaite (1995), Cole *et al.* (1998)

other fields, which of course depends on how critical one is on the supposed achievements of economics, the fact that such a 'trade' between economics and other social sciences has taken place is beyond doubt. What remains to be seen is what this exchange will lead to, and at what price.

Some attempts to link economic growth with 'social' explanatory factors have also been carried out; however, the focus has mainly been on the impact of different *institutional settings* on economic growth[2], and the major implication has been the necessity of tailoring institutions to the particular economic structure of an economic system, if the best outcomes are to be reached. In a first approach to the issue, which makes up the current dominant approach to political economy, 'institutions' are to be understood in a rather narrow sense, as they consist of patterns of economic policies, such as the income tax rate or the ideological inclination of the policy-maker[3]. Or, they may even consist of active agents of the interaction, which is then typically depicted as a 'game' between the policy-maker and the public on the implementation of an economic policy[4]. In a second approach, instead, 'institutions' acquire a somehow wider scope, as they typically consist of the whole set of social and industrial relations that take place between various actors within an economic system. For instance, so-called structuralist economists appeal to concepts such as Fordist and Post-Fordist *regimes* as 'institutional settings' that affect the economic and social outcomes of a society[5].

But these two approaches do not exhaust the role that institutions play. In particular, in the words of Douglass North (1990), many social interactions are affected by *informal* institutions, which typically are general norms of conduct that influence the practical behaviour of individuals, rather than being themselves actors of the game or variables under their control. In other words, regularities of behaviour such as social norms, customs, rules of conduct, have a direct *motivational* effect on

---

[2] In fact, that between institutions and economic growth is one of the few links that has been proved robust to nearly any empirical analysis. See Barro and Sala-I-Martin (1995).

[3] This is what happens in so-called 'politico-economic' models, where the median voter theorem determined the 'political' outcome in a society of rational 'economic' agents. For a review, see Persson and Tabellini (2000).

[4] See for instance the archetypical model of Barro and Gordon (1983).

[5] See Boyer (1997), Dosi *et al.* (1988).

individuals, and affect social outcomes insofar as the behaviour they elicit is generally shared and put into practise. In other words, economic theory has so far dealt with institutions and the 'macro'-level, but not at the 'micro'-level.

To the best of my knowledge, the impact on economic performance of this second aspect of institutions has not been taken into account by the literature, with some exceptions such as Greif (1994; 2002), who studies the impact of a 'collectivist-oriented' as opposed to an 'individualist-oriented' type of social norms on aggregate economic outcome. However, I believe that furthering this line of enquiry is of critical importance in the account of such aspects of growth as the lack of convergence amongst countries.

In particular, the failure, or lack of, economic reforms is often blamed as the main reason for the differentials in growth rates and in the patterns of convergence. For instance, if one looks at the economic policies advocated by the so-called 'Washingon Consensus'[6], you will find a programme of economic policies that is to be applied to each developing country with no distinction. When this menu failed in such a striking way as in Argentina or other Latin America countries, the answer by those who endorsed these policies was that they had been implemented too loosely rather than inquiring whether these were sound measures relative to the social context on which they were implanted[7]. What is suggested in this thesis is that at least a part of an explanation for the causes of failure of economic reform may be found in the different type of social norms and informal institutions that characterise different societies, and that particular attention to these aspects should be paid when introducing reforms in countries where this has not already happened.

However, a reader who expects to find a self-contained and fully developed model of social norms and growth within this thesis will probably be disappointed at its end. In fact, though a sketch of such a model is attempted in the last part of the thesis, the emphasis throughout the work is more on the foundational issues that can lead to such a model, rather than on striving to manufacture a final product now. More precisely, the parts into which the thesis is divided reflect the endeavour first to

---

[6] See the remarks of the creator of the phrase 'Washington consensus': Williamson (2002).
[7] On the report of such a debate, see for instance Stiglitz (2002), or The Economist (28/09/02).

analyse, and then to bring together, the two main strands of enquiry suggested above. Thus, the first part of the thesis is devoted to the study of the relationship between social norms and individual choice, and takes on the issue of how social norms come to have motivational power on individual behaviour, and how in turn they are brought about by interactions amongst individuals that affect the aggregate outcomes. The second part of the thesis develops a model of growth, in which lock-ins to sub-optimal outcomes occur because of 'co-ordination failures' between agents about the choices of the skills by employees and technologies by employers. Finally, the third part of the thesis provides some progress towards a framework in which social norms and economic outcomes can be jointly analysed.

The first part of the thesis starts off by analysing the underpinnings of a theory of individual choice, and aims to offer a background to the idea that social norms, as well as a wide range of motivations different from mere self-interest, may affect individual motivations in practical decision-making. The way by which this result is achieved heavily relies on the debate among philosophers about rationality and morality in practical choice, which also interested foundational economists such as Bacharach (1999) and Sugden (2000). The reliance on such contributions is justified by the too narrow a focus that economists have so far taken on the subject, as they typically limit themselves to rely on self-interested motivations, despite the fact that nothing 'formally' prevents them from enlarging the scope of individual motivations.

The main idea that one can draw from this debate, which emerges very clearly in the Michael Smith's book *The Moral Problem* (1995), is that when an agent makes a decision, she typically weighs up a variety of possibly conflicting *reasons to action.* These derive from the *prescriptions* that different *principles of assessment* imply in particular situations. In other words, an agent can be said to have a reason to action to do X, if this would be the *best* action in terms of the principle of assessment W, i.e. if it was an action – even though not necessarily the only one - that fulfilled W to the highest degree. Self-interest could indeed be one of such principle of assessment, but alongside it other principles such as a particular moral doctrine, an ideology, love for another person, or even whimsical desires, could provide the individual with other standards of judgement.

Such principles can be thought of as *norms*, in the sense that they offer *internally consistent* standards of assessment of the various states of affairs faced by the agent. For 'internal consistency' one can even mean the typical formal requirements that are imposed in economics for a choice to be 'rational', or even to be representable through a utility function. What matters is that the agents can rely on such principles in order to assess social outcomes in their practical decision-making without systematically incurring contradicting 'verdicts'. In particular, although morality can well offer such a principle of assessment, at this stage the term 'norm' is devoid of any feature of 'cogency' to do a particular action, which is typically associated with moral prescriptions. In fact, it only bears with it the idea of being a consistent 'metric' over the states of affairs.

Chapter 1 illustrates these concepts and sketches the debate behind them. After distinguishing between *objective* and *subjective* theories of values, the notion of a reason to action is presented. Then, the problem becomes that of assessing the extent to which the agent is able to *compare* different reasons to action. This brings to centre stage the debate on *commensurability* and *comparability* of values, which is reported with the main purpose of being aware of these issues, although in the remainder of the thesis full comparability and commensurability is assumed. Therefore, once the overall *system of ends* for the agent has been established, which potentially comprises many, possibly conflicting, norms of assessment of the same situation, the notion of *rationality* of her action can be applied in relation with this comprehensive system of ends. The final section of Chapter 1 expands on the notion of 'intersubjectivity' of values, which will prove a necessary concept when dealing with issue of how different agents come to have a shared notion of values.

Chapter 2 draws on this examination and on a review of the main empirical findings in behavioural economics to build a model of choice amenable to standard economic analysis. Its final aim is to define a *comprehensive* utility function in which the different reasons to action are represented as the various components of the utility function, and these are weighed by means of coefficients that express the relative 'value' that the agent attaches to each of them. In particular, two main types of reasons

to action will be taken into account, which, relying on related works in the literature, translate into 'self-regarding' and 'other-regarding' sources of utility.

The purpose of this exercise is twofold: on the one hand, it offers a general framework that makes it possible to deal with multiple reasons to actions in a formal way, which relies on the analytical tools of psychological games. This is shown to generalise the specifications that so far have been put forward in the literature. On the other hand, in the final part of Chapter 2, I develop a particular specification of this model that, in my view, offers a better account of the way in which individuals mix 'intentions-based' motivations with their 'social preferences'. It is argued that this specification provides a more comprehensive account of the existing experimental evidence than existing theoretical models. Its main feature is that a single individual 'reciprocates' the expected degree of commitment of other agents to the moral principle, so that a higher expected degree of compliance acts as an inventive in one's own compliance. This model builds on Rabin's seminal model (1993) of fairness, but the basic difference is that 'reciprocity' is now assessed by individuals with respect to each other's degree of compliance with the shared normative principle, rather than with each other's payoffs.

A simple application of this model to the case of a non-profit enterprise is offered at the end of the Chapter, where it is shown how this institution can be viewed as grounded on the mutual conformity of its 'founders' to a *contractarian* moral principle. This has the main feature of treating *stakeholders* external to the firm's decision-making process as actual *shareholders* of it.

In Chapter 3, the reverse side of the relationship between social norms and individual behaviour is analysed. The issue of how social norms are brought about as the result of individual interactions takes centre stage. In particular, different concepts of social norms are suggested, which differ from each other in relation with their dependence on other-regarding motivations for their sustainability over time. Thus, *mutually beneficial* conventions emerge in co-ordination games, *individually beneficial* conventions are found in standard games whose Nash equilibrium is not mutually beneficial, and *other-regarding* conventions are psychological Nash equilibria *but not*

standard Nash equilibria; in fact, compliance with them is contrary to self-interest for at least some agents in the interaction.

The question of the binding nature of social norms on individual motivation is further analysed in the last sections of Chapter 3. The concept of 'normative expectation' is investigated as the typical account in which social norms come down to have a direct motivational force on individual, i.e. without being necessarily supported by any notion of individual or *collective* rationality that their endorsement could fulfil. A distinction is then drawn between *empirical* and *causal* expectation, which will form the basis for the model developed in the successive Chapters. Chapter 3 ends with the argument that a dynamical analysis may be necessary in order to discriminate between 'stable' and 'unstable' conventions, i.e. those that self-sustain even after some 'deviant' behaviour has been introduced in the population.

Chapter 4 builds on the above analysis and provides a model in which the 'emergence' of social norms is studied within a dynamical context. The main idea is that agents are concerned with the *social* outcomes that are brought about through their other-regarding utility even during the *transition* to the possible equilibrium, i.e. during the process of the *establishment* of a convention. The framework in which this analysis is developed is that typical of evolutionary game theory. Agents drawn at random from two different populations are called to play a stage game, which is given by the classical Prisoner's Dilemma. This requires the introduction of peculiar concepts of equilibria, both in the static and the dynamical case, which accommodate for the presence of beliefs of first and second order within the individual utility function, and of *continuums* of populations.

Sugden's model of normative expectations is then contrasted with the model developed in Chapter 2. It is shown that Sugden's solution, which is subject to the criticism suggested above in terms of the lack of a causal *reason* to action, is *unstable* in a dynamical sense – or, better, stable in the Liapunov but not in the asymptotical sense - at least with respect to a strong notion of dynamical stability that does not require expectations to be consistent with the would-be equilibrium. The reason for this result is that when a norm is not supported by some *justification* in terms of either individual or inter-subjective rationality (i.e. neither self-interest nor public interest sustain the

norm) then the system does not manifest the tendency to return to the previous equilibrium when some 'mutant' agents behave in a different way from the norm. By contrast, the model developed in Chapter 3, being built on the idea that reciprocity occurs on some shared normative functions, is not subject to this criticism. However, it may still lead to some of the paradoxical results obtained in Sugden's model, which I have called *exploitative* equilibria, under some particular specifications of the normative function. In this sense, the model I propose can be seen as a generalisation of Sugden's.

The second part of the thesis develops a model of economic growth, which shares with the first part its 'evolutionary' flavour. The conditions under which multiple steady states and 'poverty traps' occur are analysed, hoping to shed some light on why development patterns differ so strikingly among rich and poor countries, and also why different 'clubs' of convergence take place within more restricted groups of homogenous countries. The model studies a multi-sector economy, where each sector is associated with a different technology whose pattern of technical change is localized; that is, technological information is a public goods at the sectoral level, but not at the aggregate level. These technologies demand workers with different skills, so that two different labour markets – one for skilled and the other for unskilled labour – exist.

Two other basic methodological assumptions differentiate this model from the 'neo-classical' tradition. Firstly, there exists a multitude of agents who are boundedly rational, thus the process of adjustment toward optimality is not instantaneous; moreover, the optimal action continuously changes as the economic system evolves. Secondly, prices are not perfectly flexible, so that markets cannot adjust to equilibrium instantaneously.

In the first version of the model, which is developed in Chapter 5, capital is the only factor to be mobile across sectors, whereas labour supply remains fixed in each segment of the labour market. Because of bounded rationality constraining the movements of firms across sectors, the share of capital invested in individual sectors follows a differential equation that is continuous in time, which is shaped as a *replicator dynamic*. The main outcome of the model is that two steady states, one carrying higher growth than the other, are both stable attractors. Hence, a result of lock-in to a

poverty trap can occur in the presence of unfavourable structural conditions in the economy. These steady states are characterised by the entire allocation of capital into one of the two techniques. Such a result of 'convergence' crucially depends on the presence of increasing returns to scale at the sectoral level, which is caused by localized technical change. Chapter 6 checks whether this result is due to the presence of a rigid labour supply, thus it allows for movement of labour supply across sectors, up to some adjustment costs. However, it is shown how the main results obtained earlier are robust to this generalisation.

This multiple steady states result is interpreted as implying that poverty traps can occur for causes that differ from those usually stressed in the literature. What we witness in this model is a peculiar *market failure* in that market forces do not provide enough incentives to individuals in order to *co-ordinate* on the efficient technique. In particular, this happens when the economy is affected by particularly adverse *structural* conditions, e.g. a relatively high skill shortage in the no labour mobility case, or high skill upgrade costs in the mobility one.

The third part of the thesis aims to bring together the two lines of enquiry developed in the two previous parts, by introducing a model of social norms and economic performance. Individuals are supposed to behave in accordance with the model of choice developed in the first part, whereas the model of the second part is used as a general framework to account for the emergence of multiple steady states, and poverty traps in particular, within co-ordination types of interaction.

More precisely, Chapter 7 develops a model of horizontal integration on the lines of Grossman and Hart's seminal model of the firm (1986), which is intended to address the question of the relationship between social norms, technology choice, and uncertainty. In particular, a distinction is made between co-operative and competitive *economic activities*, which have clear-cut interpretations in terms of the contributions required in the production functions, and *social norms* favouring competitive or co-operative behaviour. This depends on the way *uncertainty* is distributed amongst individuals, i.e. if uncertainty is managed at a collective level or if it is borne by individuals themselves. The main idea is that there exists a 'complementary' relationship between social norms and economic activities, in the sense that more

efficient economic activities favour the establishment of the social norms that are functional to them, and that, in turn, social norms can also play a role in the determination of the efficiency of economic activities.

A model is developed that aims to capture in a stylised way these considerations. Two agents have a choice of whether to invest in *co-operative* as opposed to *competitive* skills at the first stage of the game. Such a choice of the 'type' of human capital affects the probability with which co-operative and competitive *individual* productivities are determined at the second stage by Nature. Quite obviously, a higher investment in either type of skill determines a higher probability of drawing high productivity in the related option. Then, agents observe only their own pair of productivities, and decide whether to *race* for the market against the other agent, or to settle down and *co-operate* in a joint production function with the other agent. Clearly, the higher individual *competitive* productivity is, the higher the probability of winning the race, if this option is selected by at least one agent. Also, the higher *co-operative* individual productivity is, the higher the chance that *joint* productivity is also high, in the case that *both* agents opt to co-operate. In particular, two cases of complementarity and substitutability of individual productivity within the joint production function are investigated.

Numerical analysis helps us to find out that the situation behind the game can be reconstructed to very simple 'normal form' types of interaction, such as the Prisoner's Dilemma, a symmetric co-ordination game with an equilibrium that Pareto-dominates the other, or a hawk-dove game. Which scenario will occur depends on the characteristics of the two technologies associated with co-operative and competitive activity, respectively. In particular, a Prisoner's Dilemma scenario can occur when the co-operative technology is Pareto-superior to the other, but the inherent risk associated with investing in co-operative skills makes this a dominated strategy. It is argued that in this circumstance co-operative social norms could have the function of eliciting the sharing of risk at the aggregate, collective level. The analysis of the first part of the thesis shows how such social norms can indeed arise. In the case of the co-ordination game, the analysis of the second part makes clear how both equilibria, even the inefficient one, can arise as steady states of the dynamical system.

Moreover, the effect of individual risk-aversion is investigated. It is shown how this may create a peculiar type of inefficiency: agents perform the less risky activity at the first stage, which consists of investing in competitive skills, and then perform again the less risky activity, which is now co-operating. This implies that the skills they have developed at the first stage are not well suited to the requirements of technology, making the competitive technology the Pareto-superior one. In this case, relying on the same intuition as before, social norms favouring competition are likely to emerge.

The relationship between social norms and technology is finally discussed: it is argued that social norms can sometimes have a 'progressive' role, in that they act as a form of 'social capital' helping to select the best economic alternative, but sometimes can also take on a 'conservative' role in preventing the society from switching to better technologies when they become available. In the concluding Chapter 8, it is argued that this analysis may shed some light on the underlying causes of the institutional and social reform processes that are taking place in many countries.

As will become clear, a thread that links the various contributions together is the dynamical perspective that is assumed throughout. This is something more than a mere 'technical' assumption: it is motivated by the belief that dynamics is indeed a key part in the account of social and economic facts, so that something substantial would be lost by only relying on a static investigation. More precisely, a dynamical perspective enables us to look at the 'transition' of social and economic systems along their convergence to equilibria, if any, thus enabling us to spell out the conditions upon which these are selected. But there is an even more fundamental aspect: applying a dynamical analysis is functional to the bounded rationality approach that will be adopted. In particular, replicator dynamics will be seen as a model resulting from a process of slow diffusion of information amongst the interacting agents, which seems a sensible assumption in a large number of socio-economic circumstances. It is thus hoped that the models developed in this thesis aid to constitute a useful framework for a comprehensive analysis of economic growth and social norms, as well as for different economic problems calling for a dynamical analysis.

FIRST PART

NORMS, INDIVIDUAL CHOICE AND SOCIAL OUTCOMES

INTRODUCTION

That social norms have a prominent role in shaping individual behaviour is a common assumption in many social sciences, and something of which economists and, more generally, students of Rational Choice using the economist's methodology, are becoming more aware. That social norms are ultimately the result of individual interactions is something that economists are probably more familiar with than other social scientists, but on which some progress has been made by them only recently. It is the purpose of this chapter to explore the two aspects of the relationship between social norms and individual behaviour, and to advance some suggestions within this debate.

What I perceive to be the main limitation in the growing body of literature on the subject is that the two aspects of the relationship are dealt with separately, as I will try to explain briefly in this introduction. To be sure, breaking down the study of an intricate matter into its main components may be seen as a necessary step in order to obtain further development. However, since significant, though not entirely satisfactory, progresses have been made in either field, the time seems to have come for an attempt to link explicitly the two lines of enquiry. Moreover, the habit to look at only one side of the coin may cause a distorted view of the entire matter.

Let me first explore the state-of-the-art on the first side of the relationship, going from individual behaviour to social norms. Needless to say, the need to ground aggregate phenomena in individual behaviour is one of the main tenets of Economics, the so-called methodological individualism. Therefore, an account of social norms that aims to be consistent with this epistemological principle should not simply take social

norms as granted, but need to show how they can be the result of interactions between individuals, who, moreover, are supposed to take rational decisions. It is true, in principle, that this may look like a daunting task, since, by definition, the context to which social norms apply is that of a relatively large and possibly heterogeneous population. Conversely, Economics has traditionally dealt with the simplest types of situations, involving either parametric choice – namely, an individual facing a problem of choice among a well-defined set of alternatives, with some probability distributions over the correspondence between actions and outcomes – or strategic interactions within limited groups of agents. However, it is also true that for many purposes the consideration of two-person games may be sufficient in order to get basic insights into some types of interactions. Consequently, one can identify a social norm as a Nash equilibrium within this context: this is the route taken by David Lewis (1969), who finds how the game theoretical definition of Nash equilibrium fits well with his philosophical definition of convention – a regularity of behaviour among a population that creates a consistent and concordant system of mutual expectations. Moreover, the recent application of the evolutionary paradigm to Social Sciences has added further insight on this approach, making it explicit how the pairs of agents involved in the game can be thought of as representatives of large groups of agents, who are repeatedly involved in the interaction with members of the other group. This has made it possible to refine the analysis a great deal, since the evolution of aggregate outcome and its interaction with individual decision-making can be treated within the same framework, at the cost, admittedly, of very rudimentary decision rules and somehow restrictive assumptions on how information spreads through the system.

Therefore, far from saying that all the related problems are solved, in principle the economist's tool-kit appears to be suitable to make the problem of deriving social norms from individual behaviour a treatable one. However, within this account the aspect of how social norms influence individual behaviour is, in my view, limited. Norms are depicted as cognitive tools to solve co-ordination problem among the number of equilibria that could arise in a game. Therefore, norms would only act on the side of the beliefs that agents have about others' behaviour. However, this view has been challenged in that another, possibly more important, role of norms within

the individual system of deliberation is neglected – namely, the motivational force of norms of behaviour. That norms must take on this function has been debated extensively, and significant support and "empirical" evidence has been gathered around this idea. In particular, a large amount of data has been collected within experimental economics, and the idea that norms play a considerable motivational part in people's actual behaviour is now widespread. For instance, norms of fairness seem to be relevant for many people involved in some form of a division game; norms of reciprocity appear significant in contexts of repeated interaction, and so on. What is most notable, surely, is not just that people follow rules of behaviour somehow codified and become a routine, but that this goes against their self-interest, at least in the 'experimental' environment in which those observations are carried out.

Such a violation of the prescriptions of rational choice theory calls for some amendments. One route followed by some scholars is to say that experimental evidence is not conclusive, because of its methodological shortcomings, and that in any case a departure from standard theory of rational choice would leave the scholar with no viable model of individual behaviour. On the other hand, other scholars try to extend the range of motivations that individuals would take into account when making a decision, attaching in some way a role to social norms. It is this second route that I would like to pursue in this contribution, which leads to the second aspect of the two-sided relationship between norms and individual behaviour I have begun with. In fact, the standard framework of rational choice can be easily extended in order to consider a variety of motivations for the agent, which includes self-interest as well as other motives to action: what matters in order for a choice to be called rational is the *consistency* among the choices that, at least ideally, can be made, rather than whether these satisfy an individual's self-interest. If rationality has come to be closely identified with self-interest by many authors, this is the result of a restrictive, possibly myopic, use of the theory, not of its necessary implications. In other words, once the set of ends has been set out clearly, the standard requirements that the received theory of rational choice prescribe can be called upon in the same way as before.

Admittedly, this is all but the end of the story; but it should also be a starting point for a new endeavour. The set of other types of motivations that the individuals would

embrace in addition to self-interest is vague. Some scholars argue that there is the concrete risk of finding ad hoc motivations for any particular situations we study, thus making the enterprise of constructing a *general* theory of individual behaviour a utopian task. However, I believe that this objection should not discourage the economist from further exploring this approach. In fact, identifying social norms as relevant for individual behaviour, even though their exact influence should be carefully determined each time, is nonetheless a step forward in the theory; it goes without saying that many times self-interest appears a vague concept too, but this does not prevent students from adopting it as a prominent tool in their analysis without specifying its content.

Indeed, there have already been some attempts to model how social norms enter individual preferences, often elicited by the need to account for new pieces of experimental evidence. Probably, the first author to deal with this matter has been Adam Smith (1759), and in recent times John Harsanyi (1969), who identified the two main motivations of individual behaviour with economic gain and social acceptance. Since then, many authors have argued on the same lines, by trying to add to self-interest a second order of motivations including some relations with social norms. In broad terms, these can be grouped into theories of normative expectations, in which people care about the judgement of approval or disapproval that other members of the community make on them; theories of 'social preferences', where individuals form preferences on the overall distributions of payoffs amongst the agents as well as on their particular share; and, finally, theories of 'intension-based reciprocity', where individuals 'reciprocate' the intentions of other agents' behaviour by an action of the same sign.

Despite the fact that all of those attempts are consistent with the inclusion of social norms into the individual system of choice, what matters is that only very rarely has the analysis tried to link this aspect with the other side of the problem that I have emphasised before; in other words, in all those specifications, a social norm seems to be already established and taken for granted by the agents, thus somehow overlooking the issue that the norm itself has sprung out of individual interactions.

In what follows I would like, firstly, to review in more detail both aspects of the problem that I have identified. Secondly, I hope to offer some contributions on how

the two-sided relationship between norms and individual behaviour can be tackled in a unified framework. The key for this task will be to adopt a dynamical analysis, in which, on the one hand, social norms are seen as equilibria of a game, or as steady states of an evolutionary game. On the other hand, people will have some disposition to act in accordance with social norms entrenched in their system of motivations, even *during* the transition to the equilibrium. It will become clear that one relevant issue will be how agents form expectations *outside* equilibrium, which, of course, is *the* real foundational issue that all the theory of rational choice is striving to answer. I will be content with offering some insights concerning the topic, and showing some results when some particular models of learning are considered.

The first three Chapters deal with the issues raised before separately, but in reverse order with respect to the present introduction. So, the first part is devoted to an analysis of the foundational issues regarding individual rational choice, aiming to offer a model of how social norms can become internalised within individual motivations. First, I seek to clarify the philosophical underpinnings of a system of choice based on a variety of *reasons to action*, in which both self-interested and other-regarding prompts to action are included. The central problem in this discussion is the identification of the source of *value* that individuals attribute to the outcomes of their choice, to which objective standard, or *norms*, of assessment, can be associated (section 1.1). Reasons to actions can be said to exist whenever there is a source of value, and a related norm of assessment, supporting that action. Even though this approach appears to be consistent only with an objective theory of value, I try to explain how it can be made coherent with a subjective theory as well, such as the one I wish to pursue (sec. 1.2). Closely associated with the issue of the existence of different reasons to action is the problem of their commensurability, which will be analysed in section 1.3. The issue of intersubjectivity is also briefly analysed in section 1.4.

Chapter 2 aims at developing a simple, but general, representation of a utility function embodying both self-interested and other-regarding components; a number of models that have been recently put forward in the literature are illustrated as particular specifications of that function. In section 2.4 I put forward a model of individual choice that takes into account some of the criticism that I will have raised in

the previous parts. In this model I try to combine in a single framework the three approaches sketched earlier, employing the analytical tools of Psychological Games. Other-regarding behaviour is here characterised as a conditional willingness to comply with a normative principle, and the aspect of mutual expectations is obviously relevant as well. A simple application of the theory to the case of the constitution of the non-profit enterprise is provided in section 2.5.

Chapter 3 deals with the other side of the coin – namely, modelling social norms as a result of individual interactions, and builds on the distinction between self-regarding and other-regarding motivations illustrated in the first part. In fact, depending on the type of commitment required of the agents in upholding a norm, and in particular on the extent to which other-regarding motivations are to be called upon in order to sustain it, three different notions of social norm can be distinguished. First, *mutually beneficial* conventions are the typical upshot of co-ordination games (section 3.1); second, *individually beneficial* conventions lack the property of mutual maximisation of one another's payoff (section 3.2); finally, *other regarding* conventions call for agents going against their own self-interest, and are endorsed on the grounds of other-regarding motivations (section 3.3). Each notion corresponds to a different concept of equilibrium, namely *mutually beneficial* Nash equilibria, standard Nash equilibria lacking the property of mutual benefit, and finally *Psychological* Nash equilibria. The different cognitive structures that are needed in order to generate their self-perpetuating character will be analysed in section 3.4, and in particular the lack of an account for how the net of agents' expectations develops along the process of "convergence" towards the equilibrium is recognised as one of the main shortcomings of the theory. It is thereby argued that the aspect of convergence of expectations toward a stable pattern is crucial for a thorough understanding of the nature of norms (section 3.5).

Finally, in Chapter 4 such a model of individual choice is supported by a dynamical analysis in order to attempt to depict a situation where norms and individual behaviour are jointly considered and evolve together. In particular, I investigate whether some Psychological Nash equilibria, which, as set out before, support the widest category of social norms that I have identified, are consistent with a process of expectations formation converging to that equilibrium. In other words, I test whether

a social norm is 'stable' to a perturbation of the strategies and system of expectations supporting a particular equilibrium. In order to do this, two different notions of steady state equilibrium and stability are used (section4.3), one based on the original work of Geneakoplos *et al.* (1989), the other relying upon Van Kolpin's refinements of this concept (1992). Such notions are used to assess analytically the criticism put forward in Chapter 3. The relative analysis shows that normative expectations theories are characterised by a peculiar 'conformative' character in that they tend to perpetuate the current position. In this sense, such theories are interpreted as accounts of the *ex-post* sustainability of social norms, but they cannot rely upon to explain their emergence. Finally, the underpinnings of the theory based on the conditional compliance to public interest are further illustrated through a comparison with the normative expectations theory. In section 4.4 the dynamical apparatus previously developed is applied to the latter theory, and it is shown in which sense it generalises Sugden's model of normative expectations.

CHAPTER 1

FOUNDATIONAL ASPECTS OF RATIONAL CHOICE

1.1 VALUES AND DESIRES

1.1.1 Objective Vs. Subjective and Relative Vs. Non-Relative Values

Before starting off in defining the notion of value, we need a simple account of a typical situation of choice and the definition of the related terminology. I will here follow the approach common to most game theoretical accounts of a situation of choice (for instance, Myerson (1991: Ch. 1)). A situation of choice can ideally be seen as being made up by three elements: actions, outcomes and beliefs. The set of *actions* that an agent can undertake represents the way in which some states of the world are affected by the agent's action, or the probability with which some states of the world are realized. I call *outcomes* the set of possible states of world that can be affected by the agent's actions. In situations of perfect information, the relation between actions and outcomes is merely *physical*; that is, the impact that an agent's action has on the set of outcomes can be, in principle, objectively defined. In situations of imperfect information, the agent may be unaware of the exact relationship between her[1] actions and the final outcomes, thus she will form *beliefs* about this relationship; these are a set of probability distributions over the outcomes, which are conditional on the action taken by the agent. In other words, under a certain belief b_i, the agent holds that the

[1] Throughout the thesis, the subject of a sentence is always expressed as a feminine noun, whereas the second party of the interaction is a masculine noun. In particular, in the frequent two-person relationship that will the subject of analysis, the active agent will always be intended as a 'she' and her counterpart as a he. Besides, when it is the case, the third party of an interaction will always be referred to as a female. Finally, people who perform nasty or sly actions are generally referred to as males.

states of the world will be carried out with a certain probability distribution whenever she uses action a_j. Beliefs can also be thought of as indicating the initial states of affairs that the agents think of as true.

After the relationship between actions and outcomes has been defined, and the agent is aware of it through her system of beliefs, the notion of *value* comes into play. This permits to order the outcomes in relation with the degree of value they involve, so that a relation of *betterness* amongst the outcomes of the choice[2] can be created (Broome (1999: Ch 10)). Hence, a system of *preferences* over the outcomes can be generated such that it simply reflects the relationship of 'betterness' previously defined. Such a stylized model of choice can also be restated with a different terminology, which is also common within the literature. In fact, the value attached to the final outcomes by the agent can be said to be generated by what her *ends* are, whilst the set of feasible action may be thought of as the number of *means* that are available to the agent. In this way, the relationship between actions and values corresponds to that between means and ends.

This brief discussion immediately raises the question of *what* value actually is in such a construction. It is at this point that the distinction between an *objective* and a *subjective* account of value becomes relevant. According to the former approach, value is objective in that there exists a *standard*, or *norm*, of assessment of the outcomes that can be defined in naturalistic terms. In other words, the extent to which a state of the world satisfies the characteristics that are relevant for the notion of value, can be 'objectively' measured on some scale. The existence of an independent standard, or norm, of value urges the agent to order the outcomes of her choice to respect the prescriptions of such a standard; hence, once the notion of value is known, the value attributed by the agent to the possible outcomes of her choice may be objectively determined. The most common, at least for historical reasons, example of an objective account of value is provided by the utilitarian theory of choice: in broad terms, value here consists of people's happiness, which can be associated with psychological states

[2] In order to do so, we have to assume that outcomes contain all the elements to which agents attach value. For a more extensive discussion of this aspect, see section 1.2.2.B.

of mind like pleasure and satisfaction[3]. These can be, at least in principle, measured in objective terms and used as a standard to compare the value of a state of the world for different individuals. Others examples of objective standards determining values are self-interest, or prudence.

The subjective theory of value overturns the terms of analysis with respect to the objective one. Rather than speaking of some objective notions of value generating a standard for assessing the outcomes, students upholding this approach prefer to *take as given* the relation of preferences between the states of the world that an individual actually has, and to *derive* the notion of value from that. In other words, an individual's preferences over outcomes, for the mere fact to be *her own* way of settling her business, generate a subordinated notion of value. On this account, value consists of the fulfilment of our desires (Gauthier (1986: Ch. 2)). Notice that it is possible to argue that we have different *strength of desires* (Griffin (1991: 55)) over the outcome of our actions, thus making it possible to generate an *ordering* of our preferences. Therefore, on this account, value can be thought of as the *degree* to which the preferences of an individual can be said to be fulfilled by a choice. In other words, individuals can be thought of as having *desires*, *needs,* or *wants,* and that an action carries a value insofar as it helps to satisfy them. The most notable example of this approach (at least for economists!) is the theory of revealed preferences, where the agent is supposed, at least in ideal terms, to have expressed her preference between each pair of feasible alternatives in a consistent way, the utility function acting as a formal device to represent such ordering.

James Griffin summarises this discussion by saying that if desires are *subordinated* to an independent conception of value, then we are in front of an *objective* theory of value, whereas in the opposite case, where desires determine value, we are dealing with a subjective theory (Griffin (1991); (1986: Ch 3)). In fact, these theories represent only the extremes of a *continuum* of models of choice, where, as we shall see in the next

---

[3] Broome (1999: Ch. 1) stresses how there is a fundamental ambiguity in Economics about the use of the term 'utility'. In fact, this notion originally was associated with an objective account of value: in Bentham's first definition, utility was given by the "usefulness" of an object, that is by its *tendency to produce good.* It was only after Robbins that the term utility turned to be considered a measure of value, and not value itself.

sections, in the middle lie models in which the agent uses, to some extent, 'objective' standards to review and redesign her own desires.

Gauthier points out that the main difference between an objective and a subjective theory of value lies in the *locus* where value stems from (Gauthier (1986: Ch. 2)). To endorse an objective conception of value is to believe that a certain notion of value is inherently present in the world and in the states of affairs, independently of what agents' thoughts and desires are. Accordingly, we may describe value as an ontological feature of the universe, in line of principle accessible to human knowledge in the same way as the natural features of the world can be discovered by scientific or speculative inquiry. We might thus conceive of 'objective moral facts' in exactly the same way as we think of natural facts, or, if we want to deny such strict an analogy, we could say that the access to the world of moral judgements is all the same guaranteed by way of intuition (Ayer (1936)).

On the other hand, in a subjective account of value we believe that value does not lie in the world outside, but is rooted in the individual herself. Of course the individual is affected by the external world, and thus in forming her own system of values she will be influenced by that. Nevertheless, the source of value lies in the affection that the agent receives from outside, not in the object outside affecting her. Notice that an objective conception of value would not deny the existence of such a relation of affection between the outside and the inside of the agent - after all an individual must internalise her mechanism of choice, even if this stems from the outside. But the objective conception would attribute value to the object affecting the agent, whereas a subjective account would point at the affection itself as source of value.

Gauthier also points out that the dichotomy between an objective and a subjective conception of value is usually coupled with a distinction between relative and absolute values. Relativism is the notion for which value is dependent on each individual's own affective relationship with the external world. As a consequence, each person will build her own conception of the good, independently from that of others. Conversely, in an absolutist conception, values will be the same for all the people, thus shaping an idea of common good for the whole of the community. The difference between the two conceptions lies in the relation between the good *for* one person and the

straightforward good. On the first conception, if my neighbour experiences a better situation, then I am better off only insofar his good enters into my system of value - i.e. only if I am to some extent altruist and my neighbour's well-being is valuable to me. On the absolutist conception, the good of my neighbour is automatically transmitted to the whole of the society, and thus to myself as well; if improving the condition of my neighbour *is* good, then, *by definition*, that must be good for me as well.

Whilst it is natural to associate a subjective conception of value to a relativistic one and an objective to an absolute one, the other alternatives are more difficult to be tackled. Classic utilitarianism is the clearest example of a moral theory that aims to be subjective, as value is grounded on individuals' preferences, and absolute, since everyone shares the idea that the good is the "greatest happiness for the greater number". However, it is well-known that Mill's attempt to prove the validity of the utilitarian principle on the grounds of a subjective notion of value leaves many critics unsatisfied. Likewise, an account of value that was objective and relative seems inconsistent (Gauthier (1986: 53)). Such a distinction between relative and non-relative accounts of value has been largely disputed by philosophers. While Gauthier and Williams (1972: 20 ff.) endorse a relative notion of value, in accordance to their subjectivist approach, Smith (1995: Ch. 6) opts for a non-relative account. In the next section I shall try to clarify how the different theories of values affect theories of choice, also trying to depict the features of the 'intermediate' positions.

1.1.2 From the Taste Model to the Perception Model

According to Griffin (1991: 385), two different models of choice obtain when we embrace the two theories of value depicted above: the Taste Model and the Perception Model, derived from the subjective and the objective account of value respectively. Griffin stresses how the models employ the usual separation between a rational side of human nature (judgement, understanding, perception) from an attitudinal side (feeling, sentiment, will), but differ in the emphasis put upon them.

1.1.2.A The Taste model: the Pure Humean Approach and the Coherence-and-Efficiency model

The Taste Model draws on the Humean approach to human psychology. On that account, the only role of rationality is to find the best means to bring about the desires that the agent has. Desires are not subject to any activity of rational deliberation; that is, we cannot argue that it is unreasonable to hold a certain desire, or that our well-being would be improved if we assumed a new desire; in the words of David Hume, "*rationality is slave to passion*". In order to find the best means to realise a certain desire, the agent will draw on a certain pattern of beliefs about this relation. Beliefs are conceived as distinct elements from desires. Accordingly, while a belief is subject to a cognitive judgement in terms of its truth/falsity, desire is not. To be sure, some desires are directly formed on the basis of a belief, thus if the related belief is false we can argue that the corresponding desire is grounded on a wrong basis. But, apart from this case, desires are never questionable.

The Taste Model provides a somewhat implausible account for a theory of rational choice. Even if we want to leave aside the question of the content of desires, so that the *"destruction of the world can be preferred to the scratching of my finger"*, like in the well-known Hume's paradox, it has been argued that at least a minimal requirement of coherence between the choices carried out by the agent must be provided. This leads us to a refinement of the pure Humean model, which can be called coherence-and-efficiency model (Hill (1997)). This also represents the standard account of rational choice employed in Economics, which is consistent with the Bayesian approach to choice. What has been argued is that a choice should at least fulfil some basic requirements of internal coherence, without which an individual would be exposed to potentially infinite losses, as in the so-called money-pump argument – e.g. Cubitt and Sugden (2001). The more stringent condition that is thus imposed on a system of preferences in order to be considered 'rational' is that of its transitivity. Other conditions that are imposed have arguably a more 'technical' character, such as that of reflexivity and completeness.

The requirements mentioned above provide the representation of individual preferences by way of an *ordinal* measure, whereas the introduction of more restrictive

assumptions - namely, continuity and monotonicity of preferences - allow the constitution of a *cardinal* measure. In this setting, an agent's utility is simply a *measure* of value, not its source. This is coherent with the subjectivist approach, for which value lies only in the preferences of the agents, and the utility function permits a synthetic representation of them.

1.1.2.B A Moderate Humean Model

The coherence-and-efficiency model seems to offer a rationalisation of the Humean model leaving intact the basic features of its approach. Is such a model adequate to offer a comprehensive account of rational choice? Many commentators question the plausibility of such a model, even after the introduction of the logical refinements. Firstly, even the imposition of the transitivity requirement does not rule out the possibility that the agent holds blatantly counterintuitive patterns of preferences, which could hardly be judged as 'rational'.

Further, Broome (1999: Ch. 5) has pointed out that the transitivity requirement risks to become a void concept if a modification of the usual approach to the examination of preferences is introduced, thus becoming an ineffective instrument to restrain our preferences in practical terms. In fact, so Broome argues, human psychology may be such that the same option can receive a different evaluation by the individual in relation with the option with which it is being compared, and which would be turned down if the first option were chosen For instance, if the individual is comparing the three options on how to spend her weekend of 'climbing a mountain', 'going to the sea', and 'watching TV', then the option 'going to the mountain' spawns the two 'finer' alternatives of 'going to the mountain *and* not watching TV' along with 'going to the mountain *and* not going to the sea', depending on the alternative option with which it is being confronted. The set of 'finer' options, then, includes six alternatives that stem out of the original three. As a result, every possible ordering under the 'coarse' alternatives that violated the transitivity assumption could be redefined in terms of the fine alternatives in such a way that such an axiom would not

be contradicted[4]. In general terms, what is here criticised by Broome is the reliance on the assumption of independence from irrelevant alternatives, whose omission makes it possible to re-define alternatives in the 'fine' sense. Broome's conclusion is that if alternatives are allowed for, and this does not seem too strong an assumption at the light of our insight into human psychology, then it is always possible to formally justify every set of preferences, even those that at a first exam would not pass the transitivity test. Hence, the position of what Broome calls 'the moderate Humean' in fact collapses to that of the pure Humean, thus attracting the same type of criticisms illustrated above.

Finally, Griffin (1991) emphasises the inadequacy of the Taste Model, and in general of every model that relies upon a subjective theory of value, in offering a sensible account of *moral* preferences, at least within the Kantian approach. Drawing upon the Kant's theory on the subject, he points out how morality belongs to the sphere of *autonomy*, thus requiring the independence of the judgement from the sensible world. Therefore, he endorses what he calls a *fragmentation of utility*. Although the Taste model - or one of its refinement - might be tolerated in the account of individual choice, this would be at odds with normative analysis: after all, it is blatantly unfair to assign larger resources to the amateur of champagne than to the modest estimator of potatoes. Therefore, Griffin endorses the use of a notion of preferences based on objective values when involved in normative analysis.

A partial answer to these criticisms comes from those scholars who propose a further amendment to the Taste model, endorsing a more radical possibility for agents to review their desires. For instance, Gauthier (1986: 29 ff) suggests that agents should hold *considered, fully informed* preferences. His starting point is the observation that there exist two basic dimensions in the manifestation of preference. One is *behavioural*, and plainly consists in the actual choice that agents make. This is also the underpinning for the widespread approach in Economics called 'revealed preferences', for which the

---

[4] In the example, the apparent cycle $M \succ S \succ TV \succ M$ could be accounted for by the fact that the individual prefers going to the mountain when the alternative is going to the sea, prefers going to the sea when the alternative is watching TV, but when confronted between climbing a mountain and watching TV he goes for the last option. In terms of the finer options, then, the ordering would now

preferences individuals hold are straightforwardly what they reveal in their choice. But there is also another dimension, which Gauthier calls *attitudinal*, and that mainly lies in the expression of preferences made in speech. In fact, the two dimensions may not coincide: an agent might *declare* that she prefers not to smoke, while actually not giving up smoking. Therefore, on Gauthier's account, a preference might be held as irrational if these two dimensions do not coincide. This is an important step, insofar as it removes one of the basic postulates of the Humean approach, that is the absolute impossibility to question our desires, as manfested through choice.

Building on this point, Gauthier states that the preferences that form the utility function of the agents must satisfy three requirements: the agent must not suffer lack of relevant information, lack of experience and lack of reflection when formulating her preferences. If even one of these three conditions fails to be respected, then we should say that the agent has formed only a *tentative* preference, which will - or should - reviewed after the agent obtains full information, or experience, or reflection. Likewise, Brandt (1982) states that a desire is rational only if it survives criticism by facts - a person acknowledges all relevant facts - and logic - a person does not incur in logical errors when drawing inferences about her choices.

All of these attempts to introduce some methods to rationally review the desires lead us toward a 'less extreme' Humean model, if not to a 'quasi-objective' account of preferences. However, as Griffin (1991: 56 ff.) argues, this does not seem enough. After all this process of rational reviewing, the agent may turn out to still possess whims and idiosyncrasies that would surely embarrass a 'moderate' Humean. And, most of all, it is not entirely clear when these conditions of full information, experience, reflection may be said to be met. All of us agree that a better reflection may improve our understanding of a certain problem, and thus the adequacy of our choice. But reflection is time-consuming, gathering of information is generally costly, as well as the acquiring of experience. Clearly, there is a trade-off between the gain in investing resources to get a more reflective choice and the loss of time and money for such an investment. So, where is the balance? It is very hard to answer, and, further, if

---

be: $M_S \succ S_M \approx S_{TV} \succ TV_S \approx TV_M \succ M_{TV} \approx M_S$, where the label denotes the alternative with which the option is being compared.

there is an answer this will be entirely subjective, thus leaving the philosopher with no useful instrument to assess the rationality of holding a certain desire. In other words, the adoption of whatever standard to rationally review our preferences may turn out to be a void concept, practically ineffective to constrain our desires.

The solution that Griffin seems to propose is a further shift from the Taste model, to support something closer to an objective account of values. He claims that asserting that something is valuable implicitly means to compare it against a backdrop of general human values. Therefore, only relying upon that set of prudential values that is supposedly shared by all the people, we might hope that our analysis does not incur in the failures of the Humean models.

In my opinion, this claim has all the appearance of a withdrawal. As Griffin himself argues, we should not pretend to assess all the desires agents have, but only the overall desirability of their whole life. But, if this perspective may suffice for normative analysis, it seems indubitable that it leaves many other fields lacking an adequate theory of choice.

1.1.2.C The Perception Model

Indeed, what we have defined as 'moderate Humean' positions have some characters typical of the principal antagonist of the Taste model, i.e. the Perception model. The basic feature of such an approach is the presence of a norm or a standard that determines the value of an outcome and thus the preferable choice for the agent. As already seen, there exist three of such possible standard mainly considered by the scholars: interest, satisfaction or happiness, prudence.

What are the reasons to endorse such a model? In the previous section we encountered Griffin's pragmatic argument: it seems sensible to call upon objective values when dealing with normative analyses. Recall that John Rawls's (1971) proposal of drawing on *principal goods* as the main instrument of political economy is a direct consequence of the search for objective values when dealing with normative issues: principal goods are defined as those resources that are universally thought of as means necessary to reach our plan of lives. Likewise, the argument for which the adoption of

a perception model would help solving the complex problem of the intercomparability of the utility function is roughly similar.

But there also exist more pressing, philosophical reasons to endorse such a model. We have already emphasised how the basic characteristic of the perception model lies in the possibility of reviewing the desires we possess. Therefore, if values are in some way a matter of rational deliberation, then we must be inclined to accept a non-relative conception of value; as Smith (1995: Ch. 5) argues, by definition of rationality each agent, when *situated in the same circumstances*, necessarily comes up with the same solution to a problem. This must be true in particular when rational agents have to deliberate about their system of values; consequently, they should all converge to the same solution, that is to say to the same conception of value, at least ideally.

To put things differently, let us not question the existence of objective moral facts in the world, which is the view that Smith defends. Hence, we must expect from rational agents to converge to the same perception of reality: even if the world of the objective moral facts is not depictable in natural terms, thus presumably implying a greater effort of investigation to the agents, they will best exploit the information available to come up with an identical solution.

On Smith's account, a necessary feature of moral judgements is also their practical relevance, i.e. they must necessarily offer a reason to action to agents, absent weakness of will and other forms of irrationality. Thus, the common perception of a moral judgement must at the same time provide rational agents with the command to change their desires in order to abide by moral requirements. As a consequence, all the agents should converge to the same set of desires, if fully rational.

Notice that the holding of these two results - the objectivity of moral judgements and their practical upshot as reasons to action - would clash with the traditional Humean account of desires and beliefs as distinct elements. In fact, if the acquisition of the moral judgements is substantially a cognitive enterprise, then they will be held in terms of beliefs; on the other hand, desires are the prompt to actions. Thus, it would seem that beliefs influence desires. But the moral problem is solved by arguing that such a process of revision of desires does not take place in the intentional ground, but on the deliberative one. In other words, when agents act, their desires have already

been formed. The process of revision of desires has been precedent to the moment of the decision. David Schmidtz (1995) offers a similar explanation of why people should have a reason to change their desires, by means of his notion of reflective rationality, which consists of the ability to re-shape the set of ends we have in order to improve our well-being. Roback Morse (1997) offers a modification to the standard model in Economics, explicitly introducing in the utility function an endogenous mechanism to calibrate our "desires" and our "longings", where this term covers desires of a somewhat higher status, that is desires for "virtues" (beauty, culture, etc.)

1.1.2.D A Comparison

Needless to say, the objective account is not immune from criticism as well. The first concerns the philosophical justification of the idea of uniformity of values. In fact, this notion is grounded on the idea that 'rational' agents would all come up with the same solution when facing the same problem of choice. However, the requirements that agents be put in the same circumstances is practically impossible to be fulfilled; not only does it call for agents having the same conditions of choice in terms of the actions-outcomes relationship, but it also requires agents to have the same desires and the same beliefs. But this could only happen in ideal or hypothetical terms, and such an exercise would have a limited impact in practical decision-making. Smith himself admits that in reality this process might be very far from taking place, even if there actually exist large areas of moral consensus between the citizens of a society. Moreover, despite no account of rationality is offered in this account the concept of rationality itself is the main cornerstone of the value uniformity claim. This certainly raises some doubts as to the significance of this assertion. In section 1.4 I shall seek to offer a different route to how moral values could come to be shared within a society.

Consequently, the argument that endorses the objective idea of value is located in an ideal area, which only hypothetically could invest practical decision-making. In fact, convergence is subject to the full rationality of agents, which in turn implies the truth of all the relevant beliefs and the exactness in the method of deliberation, along with a concern for the systematic justifiability of the whole set of desires (see Williams

(1972); Smith (1995)). Obviously, the same criticism raised before in relation to the too 'timid' attempts of the moderate Humean to rationalise her desires can be restated here on an even larger scale. On the other hand, the pragmatic argument that an objective conception of value is more helpful than a subjective one in normative analysis appears indeed a sensible view, although the method to select an objective standard may induce some concerns as to its arbitrariness.

On the grounds of such considerations, it seems that a 'not too extreme' subjective account of value, i.e. some types of 'moderate-Humean' models, seems more defendable on epistemological and philosophical grounds. In fact, the objectivist model may be seen as a particular case of the subjective one, because an agent can always voluntarily choose to adopt an objective standard to form her preferences. On this view, an agent can actually act in order to pursue some objective standards such as her interest, or the maximisation of her pleasure or prudential values, though deciding to do so voluntarily.

Moreover, what I believe to be a point partly neglected in the literature is that an agent may form her own norms or standard of assessment, and then use them as subjective standards by means of which to appraise outcomes. In this sense, the agent may act 'as if' she was upholding an objective account of value, in the sense that her way of attaching value to outcomes fulfils all of the properties illustrated in the previous section, but value would be at any rate subjective as the norm stems from the agent herself. Another way to express the same concept is to say that an agent may decide her own ends, or her plans of her life, and then act 'objectively' in the way that best concurs to fulfil such ends.

Furthermore, the 'intermediate' account between the taste and the perception model could simply be offered by considering that some normative standards of assessment exist and are 'objective', in the sense that the perception theory argues for; nonetheless, the individual is free 'to choose' among them and 'combine' them in the way she most likes. To be sure, this account is not immune from criticism either; in particular, the problem of commensurability among values, which will be the topic of section 1.3, seems an obvious objection to the theory. However, I believe that this

account may offer a helpful starting point to find the right 'compromise' between the models illustrated so far.

On the grounds of this discussion, in the remainder of the thesis I shall adopt the 'intermediate' perspective just illustrated. So, I will take for granted that there exist some standard of behaviour, or norms of assessment, such as self-interest or prudence. These may be thought of as being defined 'objectively', i.e. in naturalistic terms, or they could stem from an agent's own process of speculation and deliberation. What matters is that the agent is free to choose among them when forming her set of preferences – or set of ends - and act accordingly. This account enables us to use the same categories typical of an objectivist account within a subjective framework. The model of choice developed in the next Chapter will be based on these ideas.

1.2 Reasons to Action

1.2.1 A Definition

The analysis conducted so far in terms of values and norms will now be extended in order to come close to a viable model of choice. The following argument is mainly based on Smith (1995: Ch. 5). In particular, what I want to describe is how the motivational side that we have depicted so far can be connected to that of actions. This connection is made easy by the structure of the model illustrated so far. In fact, I shall talk about *reasons to actions* that an agent has in a particular choice, where these are grounded on the sources of value that the agent has included in her system of ends. In other words, we can say that the fact that the agent attributes value to a particular principle, e.g. self-interest, prudence, etc., and that this principle provides the agent with a certain *prescription* concerning the behaviour that the agent should take, also offers a reason to action to the agent. For instance, I may claim that "it is *in my interest* to do X if I am in C" and, at the same time, assert that "it is a *moral requirement* for me to do A in C". Those two sentences may lead to different prescriptions in terms of practical behaviour, but they can contemporaneously be held by the same agent depending on the perspective that she is adopting in assessing a situation. What

distinguishes the different types of reasons to action is their being related to some general principle of assessment - namely, a *norm* - which allows us to state that A-ing in C is *desirable* or *required.* Therefore, we may say that we have a *normative reason* to A in C if there exists a general standard of assessment by means of which we can justify the action. As Smith claims, to say that someone has a normative reason to A is to say that there is some requirement that she As; that is to say, her A-ing is justified from the perspective of the system that generates such a norm (Smith (1995: p. 95)). It is important to stress that the sense in which we talk about a *normative* reason is related to the existence of a standard stating the desirability/requirement of that action, and not to its overridingness over other prescriptions, as some of the arguments of the next sections may lead us to think. To be sure, the different standards may give contrasting prescriptions in relation to a certain problem of choice: in those cases we have a case of conflict. I shall deal with those issues in the section 1.3.

Another relevant issue concerning the normative content of reasons to action regards the state of beliefs that the agent holds. In fact, on Smith's account, a normative reason can be said to be drawn from rational analysis only if the agent holds *true* beliefs over the relevant elements of the situation. In this case, we can say that the action can be *justified* by the existence of a particular *reason* prompting the individual to follow that particular action. Consequently, the reason and the action may be thought of as standing in a *causal* relation between each other. On the other hand, whenever the action that the agent intends to effect is derived from a logically coherent analysis but it is supported by a wrong belief about the means-end relationship in which she is involved, we could speak of a *motivatation* to action. In this case, it would be possible to give an *explanation* rather than a justification for such action, where such an explanation would have a teleological nature and would be grounded on the psychological states of mind that have brought about the action[5]. In other cases it is not possible to give any account of an action, either in justificatory or explanatory terms; in this case we say that the action was unreasonable *and* unmotivated.

[5] In the next Chapters, I will not refer to such a distinction between reason and motive to action, and I will consider the two as synonymous.

1.2.2 A Taxonomy of Reasons to Actions

In the present section I seek to put forward a taxonomy of different reasons to action that will prove useful in the rest of the work. As I shall illustrate in section 2.1.1, the literature abounds with different categorisations of types of behaviour, which differ more for the terminology adopted rather than for their content. However, a distinction seems natural to be put forward when dealing with individual motivations: on the one hand, we have the reasons of the self; that is to say, those concerning the individual who is acting. On the other hand, we have reasons to action that do not strictly refer to the self, but are determined through taking on a broader view and considering the interests of the other agents involved in the interaction. The broadest terminology that has been adopted in relation to this distinction is the one that calls the first type of reasons to action *self-regarding* and the second *other-regarding*. This dichotomy will be extensively analysed in Chapter 2, thus I will instead focus here on another distinction that appears to be somehow neglected in the literature, that between *consequentialist* and *deontological* reasons to action[6].

1.2.2.A Consequentialist Reasons to Action

Simply stated, reasons to action can be said to be consequentialist when they refer to the consequences of the agents' actions. Consider the situation of choice set out in section 1.1.1 within a context of strategic interaction involving many agents. Outcomes of actions are here states of affairs that can be described by the list of all pertinent variables, such as income, effort, psychological satisfaction, etc. for each agent involved in the interaction. The distinction between self-regarding and other-regarding preferences depends on whether an agent takes into account only the consequences related to her very self within this list, or also considers the consequences for other agents. In the former case we have *self-regarding* consequentialist preferences.

Copp (1997) associates this distinction with the perspective used in appraising a certain outcome. Accordingly, self-regarding reasons are reasons grounded in a

---

[6] This distinction has been suggested to me by Sacconi, and developed in Sacconi (2002) and Grimalda and Sacconi (2002).

person's own standpoint, which he calls *internal.* Conversely, other-regarding reasons stem from the adoption of some other perspective, such as that of some other agent involved in the interaction, that of the "team" of which the agent is part of (Sugden (2000)), and, at the extreme, that of the society as a whole. Copp calls this an *external* standpoint, implying that the point of view is different from that internal to the agent herself. To be sure, the agent may choose to adopt different notions of value when using a certain standpoint; for instance, one's needs, desires, wealth, all refer to an internal standpoint and originate different preferences. I shall further expand on these issues, and particularly on those related to the idea of *self-interest,* in later sections.

The range of possible sources of value increases even further when the agent takes account of the consequences of social interaction on other individuals, i.e. she adopts an external standpoint. In all of those cases I shall talk about other-regarding consequentialist preferences. In fact, this definition does not necessarily imply a benevolent disposition of the self towards other people, but only that individual preferences are affected by the outcomes of social action for other people, in whatever fashion this may occur. For instance, if I loathe my neighbour then the very fact that *my* ranking of states of affairs is (possibly partly) based on consequences referred to *him* makes this an other-regarding type of preferences. To be sure, though, other-regarding preferences are the basis to express an individual *moral* preference of a consequentialist type, namely a preference in which the interests of every agent is considered, in each outcome of the interaction. Altruistic preferences are another special case, in which the agent attaches a high, possibly exclusive, weight to the interests of other agents rather than her own.

1.2.2.B Deontological Reasons to Action

The second category of reasons to action is characterised by their non-strict consequentialist nature. I shall call them *deontological,* since they are based on some intrinsic characteristics of the agents' actions rather than merely on their consequences. In other words, agents are prompted to act in a certain way by the awareness that their actions satisfy some particular properties, somehow defined, of the action, rather than from the outcomes of their actions. For instance, the agents

may derive a spur to action in that the procedure followed in their choice has been fair and it has not violated the pre-constituted rights of any of the participants. Or, they may attach significance to the fact that a certain action fulfils some moral or ideological principle. In all of those cases, the agent can be thought of as deriving a reason to action from the adherence to a *rule* of behaviour, where this fulfils some characteristics that the agent thinks as significant. Even in this case it is possible to draw a distinction between self-regarding and other-regarding deontological reasons to action, where the former refers to the instance in which the agent only cares about one intrinsic characteristic of *her* own action, whereas in the latter the agent takes into account characteristics of both her own action and others'.

This argument may be subject to the following type of criticism; an outcome can always be defined so that it comprises every characteristic to which the agent assigns value, thus also possibly including 'deontic' properties of the patterns of actions. In other words, every element that the agent deems as relevant for her choice can be included in the description of the state of affairs, and in particular some deontological property of the action(s). This is essentially the theory advocated by Sen (1985, 2000, 2001), with particular reference to the notion of *freedom* as the significant deontic property. As Scanlon (2001) points out, this approach calls for a subjective theory of value, whereas only on an objective account can the distinction between 'consequences' and 'actions' leading to such consequences still be said to be significant and neat. In particular, the latter case corresponds to what Scanlon calls *Foundational* Consequentialism, which is consistent with classical Utilitarianism, as opposed to *Representational* Consequentialism[7], where value is subjectively determined by the agent *and* some notion of fairness pre-exists to that, so that the traditional means-ends relationship is no more than a formal construct of rational choice. Verbeek (2002) holds an even more radical position in arguing that the inclusion of the agent's concern for the fairness of the process is incompatible with any notion of Consequentialism. Sen's theory would lie in the middle between the two categories, in that he endorses a subjective account of value *but* moral values are not pre-determined: in this there would lie the properly consequentialist trait of his theory. According to

this view, the distinction between 'consequences' and 'set of actions' that leads to such consequences may appear somehow redundant. This would also undermine the present distinction between consequentialist and deontological preferences. However, I believe that this separation in any case helps to clarify the different sources of value that the agents deems as relevant, thus making this distinction significant.

1.3 The Problem of Commensurability Amongst Values

Our common life is full of examples of problematic choices: if we find a wallet with a large sum of money in a park, we are probably uncertain as to whether to take the money for ourselves or give it to the police; surely, this couple of conflicting prescriptions derive from the use of a self-regarding reason as opposed to some forms of other-regarding ones. The plot of many artistic works is simply based on the conflict between such reasons: in Romeo and Juliet the *reason d'etat* of obeying the orders of the families clashes with the reason to action derived from their own *sentiments.*

In section 1.2 I have illustrated the underpinnings of a model of choice that leaves as ample as possible the set of values that the agent may include into her system of ends. However, this perspective leaves one problem unanswered yet, namely that of their comparability or commensurability. Since this is a central question in the debate regarding rational choice theory, I will devote the present section to shed some light on this topic.

In Copp's terminology (1997), when values are adopted in problems of choice, they boil down to *verdicts* about possible ways of behaviour. For instance, when analysing a certain choice to be made, we may talk about the verdicts of morality, which are derived from the adoption of some moral standard, or verdicts of self-interest, related to the standard of identifying what action best fulfils our ends, and so on.

However, a problem arises when the agent considers two values implying different prescriptions: that is to say, verdicts may be in *conflict* with each other, and this may

---

[7] For a general introduction to some aspects of consequentialism, see Pettit (1993).

affect the consistency of our practical actions. The agent should then be able to *compare,* in some way, the two values in order to come up with a consistent decision, but of course this is generally not an easy task. Comparability is made possible by the existence of some common scale on which the 'degree of persuasiveness' of the two norms of behaviour can be weighed against each other. Students who are particularly sceptical about the possibility of comparing different values talk about their *incommensurability,* thus further emphasising the impossibility of comparing values at a purely conceptual level.

In what follows, I shall try to analyse the debate in the literature on this theme. Three positions will be distinguished: at the two extreme lie the opposite views about the commensurability of values, that is to say full commensurability or complete incomparability of our values, while in the middle we shall review positions in which the comparability among values cannot be fully guaranteed. The following analysis is intended to be mainly a survey of different positions on such a relevant theme. I will not attempt, though, to find possible lines of convergence between them, and in the remainder of the thesis I shall, very simplistically, assume full commensurability between values. Nevertheless, I thought that it was at least necessary to be aware of how restrictive this assumption could be.

1.3.1 Incommensurability

Copp's position is one of extreme scepticism about the inherent consistency of our practical rationality. His point of departure is the claim that in order to reach the coherence of our judgements, we need a general standard capable of letting us discern what is the action to be taken in each circumstance we come across. The main characteristic that such a general standard should satisfy is one of *overridingness*, that is to say the capability to settle all of the possible conflicts that may arise among our reasons to action.

To be sure, all of the general principles we illustrated in the previous sections, like the norm requiring acting in accordance to self-interest, or in accordance to morality, are candidates to such a role. However, Copp argues that the possibility of conflicts between such principles is so evident that we actually need a sort of *meta-standard* able

to settle the conflicts between these 'first-order' standards. For instance, the mere analysis of the dilemma between self-interest and morality is able to completely puzzle our system of judgements. It seems straightforward that in some cases we would act in accordance to our self-interest, in other moral reasons will be prevailing, thus denying a complete overridingness of one norm on the other. Other principles, like that of personal excellence, are also implausible as general norm of conducts. As already noted above, the argument for which the moral principles, *by definition,* should overcome the others rests on a misunderstanding. For this is nothing but a restatement of what moral principles are, restricting the class of supposed moral actions to that of the 'morals actions actually winning the match with other principles'.

Then, so the argument goes, the meta-standard should tell us just when a particular type of norm should prevail and when it should yield. But this is not an easy task: suppose we have found such a supreme standard. On what basis can we judge it as supreme? Of course we need *another* standard to assess its supremacy, unless we want the supposed meta-standard to be judged as supreme *from its own standpoint.* But this seems plainly unsatisfactory, because on such grounds *every* standpoint is supreme. Therefore the supreme meta-standard cannot be *unique,* but for this very reason it cannot be supreme. In other words, a *reductio ad infinitum* argument seems to be looming. As long as no 'first-order' standard seems effective as an overriding standard, and this seems justified by common sense, we call for a second-order standard to sort the conflicts between first-order ones. But how can we judge the prevailing 'second-order meta-standards? We would need a third-order meta-standard, and so forth. Therefore, it seems that the idea of a meta-standard is logically incoherent.

The conclusion Copp draws from such an argument is that of a profound scepticism about the possibility of the unity of practical reasons. As long as the recognition of a supreme principle leading our choices is impossible, the very idea of commensurability of our values seems impossible too. Of course in many situations, the overridingness of a particular principle over another will be evident, but conceptually the overall coherence of our practical rationality seems undermined by the absence of a general standard of judgement.

1.3.2 Full Commensurability

As we emphasised in the previous sections, the fundamental idea underlying the notion of value is the possibility of "measuring" the extent to which certain outcomes satisfy the norm, or principle, to which the particular value which is being considered refers. On the grounds of this picture, we can say that the idea of commensurability is twofold: on the one hand, arguing that values are commensurable means that *all* the outcomes may be compared on the basis of the principle to which the value refers, thus creating a complete ordering of the outcomes available to choice. Provided that the ordering satisfies some logical rules of internal coherence, it is thus possible to dispose all the outcomes under observation onto a single *scale*, which also generates a utility function. In this view, the problem of commensurabilism shifts back to one of *ordinalism* (Broome (1999: 146)).

Therefore, this first account of commensurability refers to a single, or indistinct, idea of value, questioning the possibility of an ordering of the objects based on such a value. On the other hand, the second account of commensurability refers to the comparability of *different* ideas of value, arguing that it is possible either to generate an ordering of values themselves or to synthesise all the distinct account of values in a single, general, meta-scale of value. In other words, we should be able to discover that meta-standard against which Copp argues. These two accounts of commensurability are certainly related to each other: we might say that once the second of the problems put forward, that of the commensurability between values, has been solved, then we can advance to the first one, related to the possibility of having a comprehensive scale of value.

As both Griffin (1977) and Broome (1999: Ch. 9) stress, this second problem is easily solved when we observe that in our comparison *we do not need to compare values*, but *options*, that is to say events or outcomes exactly specified as to the *amount* of value included in them. For example, while it seems impossible to compare in abstract terms the value of "free speech" with the value of "eating pizza", once we specify the amount of each of them we want to compare, an answer becomes possible. Thus, we can say that *'a small amount of free speech is better than any large amount of pizza'* (Broome (1999: 145)).

On such grounds, the answer to Copp's sceptic position about the unity of practical reason is that we do not really need a meta-principle according to which settling our conflicts between different values: once the choice is arranged in terms of *choice between options*, and not in terms of choice between values, the problem of finding a meta-principle simply becomes irrelevant. Commensurability is provided if we are able to *order* our options on a scale: more precisely, *if the values are commensurable, that means they can be measured on the same scale. To claim that values are commensurable is to say that options are ordered* (Broome (1999: p.146)).

If the second problem of commensurability is thus sorted out, we are left with the problem of finding a homogenous scale onto which arranging the values attributed to the various options. Regarding this point, the controversy between adopting an objective instead of a subjective account of value arises again. Further, the problem of commensurability among values within an *individual* scale is usually linked with that of *comparability* of values *between* people. In fact, we do need to compare different agent's position in the normative analysis of many social situations. Clearly, the choice of one type of scale on the individual ground affects the type and the extent of comparisons we can make from the inter-individual one.

Sen provides a review of the different forms of individual scales we need in relation to the type of interpersonal comparison we may do, based on the range of *transformations* of the utility function that might be attributable to an individual (Sen (1979: 191-194)). The possible forms range from mere *ordinalism*, where *all* monotonic transformations are feasible and which is sufficient to generate a coherent individual scale but not interpersonal comparisons, to *level comparability*, where the set of transformations is restricted so that to allow comparisons in the level of welfare of different individuals, in accordance to Rawls's maximin principle. Then we meet the first forms of *cardinality* of the utility functions when we adopt *interval scales*, in which only affine transformations - that is those of the form Wi^=ai+biWi - are allowed; they permit *unit comparability* when the parameter used to change the unity of the scale is the same for all individuals -that is, bi is the same for all i. In this case we could actually compare *how much* an individual values one option better than another one with the assessment of another individual. This is equivalent with saying that the unit

of measure of the value is the same for all individuals, but the *origin*, that is to say the location of the 0 in the scale, is arbitrary. We thus get *ratio scales* when the origin of the scale is not arbitrary but fixed, as only *homotetic* transformations are allowed. In this case the ratio of the absolute level of utility of the individuals - not only of the differences between levels like in the previous case - acquires meaning.

Both Broome and Griffin endorse some form of cardinal individual utility function, but differ in the account of them. Broome opts for an objective account, arguing that this induces unit comparability since the unit of measure is objectively determined. He also introduces a way to identify a non-arbitrary origin in the scale. Griffin reaches the same result of creating a ratio scale relying upon a subjective account, where value is conceived as *strength of desire*, arguing that it is practically impossible to measure the quantity of value involved in an outcome if we stick to an objective account. On such explanation, the origin of the scale would be given by the situation of indifference of the agent respect an alternative; further, by checking what is the strength of a certain desire we would come up with the "distance" of the value of an option from the origin.

1.3.3 Partial Incommensurability

There are two main lines of argument to the view that commensurability between values at the individual level and intercomparability of values between individuals is possible only up to a limited degree. One refers to Bernard Willliams's objection to the possibility to fit *all* the values on a single scale, because of the lack of a clearly definable unity of measure common to all of them (William (1972: 55 ff)). The other refers to the idea of *vagueness* in the comparison of alternatives, which undermines the possibility of ordering the options available because of the existence of large "areas" in which the judgement is ruled by a logic different from the "standard" one used to order values.

As a first remark to the possibility of full commensurability of values, we have to notice that so far we have always supposed that the value were found in the *outcomes* of actions, not also in the *means* that bring about those outcomes. This assumption of consequentialism is certainly functional to the utilitarian view, but it is far from being

unquestionable. Arguably, many deontological moral theories would not pose the problem of commensurability in these terms, or would not pose the problem at all.

Leaving aside this point, we now have to consider Bernard Williams's objections to the idea of full commensurability (Williams (1972: 99-104)). As he points out, inside utilitarianism there is an inevitable trade-off between the *minimality* of the requirement that happiness is the source of value for each individual and commensurability. In other words, if we want to carry out the 'quasi-arithmetical' operations of *comparisons* and *sum* that are predicated by the utilitarian principle, we would want the argument of the sum to be homogeneous. This is certainly the case if we restrict our idea of happiness to one of *pleasure* or *satisfaction*, but it clearly seems fairly unrealistic that we link all of our system of values to such an account. On the other hand, if we want to enlarge our meaning of *happiness*, thus including many more motives conducting to a happy life then mere pleasure, we end up having individual scales of values hardly comparable with each other. The point is that such an account of a happy life seems rather to be composed by a *set* of other values, ranging from integrity, freedom, love, and so on. Therefore, even if it were possible to create an individual ordering which took up all of this possibly conflicting motives, it would hardly seem possible that we could find in it a not controversial measure to make interpersonal comparisons.

Indeed, such a comprehensive account of happiness that Williams considers comes very close to the concept of utility as strength of desires that Griffin endorses. Therefore, Griffin's strenuous defence of commensurability would look in danger. In fact, he seems aware that the range of values that are commensurable is in some way limited. He honestly agrees with Williams that there may exist some values whose inherent nature prevents a full appreciation in terms of their capability to satisfy desires. The clearest example is that of justice: a state of the world in which a fair distribution of welfare is guaranteed, may be considered valuable *in itself* independently of the amount of satisfaction of desires that it may fulfil (Griffin (1977: 52)). Further, such values like beauty, human life, or knowledge could hardly be fully explained by means of the desire they spur in people. Therefore, the scale which he proposes to measure the value, that of strength of desires, does not seem suitable to order all the values we come across, confirming Williams's scepticism. Still, Griffin believes that

such a difficulty in dealing with these cases does not threaten the very basis of commensurability, and consequently of utilitarianism either. He claims that for instance the problematic value of justice can be accommodated in the utilitarian approach by considering it as a *constraint* to the maximisation principle. Moreover, all the other apparently puzzling values for this approach may well be "translated" in terms of strength of desire after opportune reflection.

The debate on the possible trade-off between the extension of our preferences and the idea of commensurability becomes even more acute if we look at the practical consequences to which it leads: when sketching a practical way to undertake interpersonal comparisons, Griffin comes very close to a sort of cost-benefit analysis, as he explicitly admits (Griffin (1977: 265 and ff.)). Thus, the key variable to carry out such operations turns out to be resources, like money or time. In fact, we said that on the conceptual point of view we should measure the "distance" of an option from the indifference position, or 0 point of the scale when doing comparisons. On a practical ground, this amounts to measure the *availability to pay* of an individual, and of course the measure of money is the most natural candidate to supply a standard for this. However, Williams contends that this point seems really paradoxical: it seems common sense that we cannot express some values, like the value of a life, or the value of an ancient town, merely in monetary terms.

If these arguments against full commensurability seem far from unconvincing, there is another point made by Broome (1999: Ch. 8) that further weakens its supporters. This is related to the idea of vagueness in the capacity to discriminate the relation of preference between two options. The main idea is that in some cases, especially when we consider a *continuum* of alternatives, ordered with respect to a certain feature - for instance, the range of colours going from red to yellow, or the attractiveness of a job as measured by the related income - and we want to compare them with an alternative not belonging to the continuum itself - like the colour reddish-purple, or another career - we come up with the impossibility to judge clearly what relation of preference exists, if any. Especially when we are in the middle of the continuum, either we shall not be able to say if one alternative is preferable to other, and *vice versa*, where this is different from saying that we are indifferent between the

two alternatives. This is the idea of vagueness, and it carries out that we cannot precisely locate the boundaries where the relation of vagueness becomes one of preference in one of the two ways.

Broome carries over this general argument to the topic of commensurability, arguing that in many situations our judgement between two values will be actually *vague*. This implies that not only will it be impossible to carry out comprehensive inter-individual comparisons, but also that this will be true at the individual level. In other words, the same attempt to build an ordinal utility function seems at odds with the existence of an area of vagueness in our judgement. Broome further emphasises that it is altogether arbitrary to assume, as Griffin implicitly does, that such an area is after all irrelevant because of its "limited" extension: nothing can exclude that the region of vagueness be very wide. More, he points out the strong difference existing between indifference and vagueness, in terms of consequences of our actions: vagueness is in fact subject to the *money-pump* argument, being thus at variance with the Bayesian account of rationality.

Linking this work with others by Broome, it seems that the solution he puts forward is antithetical to the Griffin's one, urging the adoption of an objective account of value when dealing with interpersonal comparisons and, more generally, with normative analysis. At the time being, he is nonetheless not clear about the contents of such an account. Still, his account of vagueness looks a powerful instrument to formalise the idea of partial incommensurability between values.

1.4 Inter-Subjectivity and the Sharing of Moral Values

Although this is not the right place to argue at length on the moral philosophers' search for their Holy Grail, i.e. a general theory of morality representing, at the same time, both a consistent general norm of assessment and a motivational prompt to action for individuals' practical behaviour, another relevant aspect that needs, at least, to be skimmed over is that concerning the issue of *convergence* of individuals towards the same set of moral principles. The practical relevance of this issue will be clearer in the next Chapter, where the evidence emerging from experimental economics will be

reviewed in section 2.2. It will be apparent how in many instances the subjects involved in experiments adopt a typical 'non-selfish' behaviour, which sometimes, although not all the times, seems to be driven by a concern for the interests of the 'group' rather than their own self-interest. This is particularly evident in Public Good Games, where the behaviour of agents who give to the public good and punish defectors seems to be driven by the concern for a group-oriented 'norm of co-operation' (Dawes and Thaler (1988)).

Existing theories of individual motivations do not seem suitable to account for this kind of facts, as they generally focus on self-based assessment of the overall allocation, such as the payoffs difference between the subject and other individuals (see section 2.4). Conversely, I shall argue that this fact can be accounted for by the idea that (some) individuals adopt an *impersonal* perspective in the assessment of the overall allocation, and act so as to endorse the resulting criterion of assessment for states of affairs. I shall also argue that this leads to individuals adopting the 'objective function' of the group, although this does not lead to the notion of 'team thinking' (section 2.4.5). This argument will form the background discussion for the model of motivations that I shall develop in section 2.5, in which agents have a conditional willingness to comply with a moral criterion stemming from the adoption of a common standard in assessing states of affairs. However, to make this account sound, it is necessary to be precise as to the extent to which such a function can be said to be 'shared' by individuals, i.e. in which sense individuals 'converge' to the same set of moral prescriptions. As will become clear, the fact that this account is 'shared' is – at least to same degree - necessary, as individuals condition their degree of commitment with the moral point of view to the expected compliance with others. That is, a common idea of a public standard of evaluation is needed in order for individuals to be able to infer correctly the nature of other agents' actions in terms of the compliance with the public standard.

In this section I shall put forward two possible explanations for this point. The first is to some extent 'tautological' within the lexicon of moral philosophy, and hinges upon the stance that moral values are objective. It is then straightforward to argue in favour of the 'convergence' of individual moral point of view to the same notion of

morality. A related view, which builds on the theories of political philosophers such as Brian Barry and the late John Rawls, is that morality is defined *ex-post* as the area of the 'overlapping consensus' between different individuals' moral ideas. Although this particular account will not be pursued in the rest of the work, it still remains an important benchmark for further development. I will instead turn to an alternative account that preserves the subjective notion of value, but stresses how moral judgments, because of the impersonal perspective that is adopted, end up having an 'inter-subjective' character, on the grounds of which individuals can be expected to 'converge' to the same public standard of assessment. It is on this latter account, which can be traced back to the work of David Hume, that I shall draw on in the rest of the thesis.

Let us start with the first account. After the discussion of section 1.1.1 it should already be clear how an objective notion of values requires individuals to *converge* to the same set of moral ideas. The realm of moral judgments is seen as a reality accessible either in 'naturalistic' or 'intuitionistic' terms[8], but in any case accessible to individual discovery in cognitive terms. Hence, according to this account, expressing a moral judgement is a matter of *belief* on a moral fact; that is, believing that 'it is right to do X' is tantamount to expressing a cognitive judgement about the truth of a fact. On this account, then, there is no distinction between cognitive and moral judgements: they can both be reconstructed as sentences to which the question whether they are true or false can be applied.

Therefore, one who believed in objective moral values would not encounter any conceptual difficulty in assuming that individuals appraised a situation using the same moral standard. For each individual would be drawing on the same source in order to form her own individual moral judgments, which could be reached by means of rational enquiry or intuitive perception. To be sure, individuals could still be insecure as to whether what they are doing is 'right' or 'wrong', or whether the situation under their observation is morally praiseworthy or not, but their indecisions, or mistakes, would be comparable to the indecision or mistake of a person who did not know the

[8] For the reconstruction and an analysis of the debate, see Smith (1995: Ch. 1)).

'right' answer in a quiz. In either case, adopting an invariable public standard of assessment valid across individuals would be a justified methodological assumption.

A view of morality that is to some extent related with that of moral realism is one that *defines* morality as the area of general, or even unanimous, consensus amongst the individuals of a society. This approach draws a distinction between *individual* and *social* views of morality, the latter being defined as the 'least common denominator' of the former. Hence, areas on which there existed disagreement amongst individuals as to their moral evaluation would be eschewed by the social criterion of evaluation of states of affairs, though individual judgments would still count as expression of equally creditable forms of assessment. The presence of different individuals views, rather than being detrimental for social cohesion, could instead be seen as an instance of social justice. This point is made very clearly by Barry (1995), who argues that the respect of different individual positions would be *per se* a moral requirement of a well-ordered society, and as such guaranteed by procedural – rather than substantive – justice. In fact, the respect of differing conceptions of the good is a minimal requirement of social justice within his theory of 'justice as impartiality'. Likewise, Rawls (1996) argues that social justice – and then, indirectly, morality - is the area of the 'overlapping consensus' between possibly conflicting claims coming from different social/ethnic/political groups within a society. On this account, hence, the convergence towards a common idea is to same extent 'tautological', as an individual in upholding the 'social' morality would simply be endorsing a moral principle to which he would comply as an individual.

The alternative account of the convergence of moral ideas amongst individuals within a society arrives at a notion of shared values and at the same time warranting the subjectivity of moral judgment. Clearly, a 'subjectivist' could not rely on the same type of confidence about the existence of moral facts *out there*, open, at least in principle, to be grasped by human inquiry, thus making the account of convergence more problematic. However, this result can be reached by relying on the view that individual moral judgments are based on a notion of *impersonality*, which refers to what is the *common viewpoint* amongst the members of a society on a certain issue. Once such an impersonal perspective is embraced, the resulting moral judgment will inevitably

have an *inter-subjective* character, which is then expected to imply that moral values will be shared between individuals to a substantially high degree. Hence, values are not shared as a consequence of being derived by an ultimate common source, namely, the world of 'moral facts' as the objective account purports, but because of a key characteristic of the way in which moral judgments are formed.

In order to reach a better understanding of this aspect, we need to turn to the work of David Hume, who in the following passage from section IX of the *Inquiry into the Principles of Moral* (1777) states this point very clearly. *"He must here,* [i.e. when expressing a moral judgement holding true for the group of agents he belongs to] *depart from his private and particular situation and must choose a point of view common to him with others; he must move some universal principle of the human frame and touch a string to which all mankind have an accord and symphony".* This passage highlights two basic features of the Humean doctrine of morality (Lecaldano (1991)), which are the universalistic nature of moral judgments, and its subjective character. I will briefly comment on these aspects in turn, and then come back to the question of how moral values can be expected to become shared.

As for the first aspect, what Hume depicts here is to be intended as a descriptive *procedure* of how human beings construct and shape their moral judgments in real life, rather than a *normative* criterion of how people *ought* to form them. In other words, on Hume's account it is a natural trait of human beings to express judgments by taking a standpoint common with other people, which can arrive to embrace the whole humanity. It is precisely the adoption of such a standpoint common of a person with others that makes for the impersonality and the universality of the judgment (Wiggins (1991: 60)).

To be sure, it still remains open to question whether such a common point of view can be relied upon to exist in most of the relevant situations in which the elaboration of a social standpoint is needed, and, most of all, whether *in reality* different individuals will come to adopt the same common point of view. However, Hume points out how our language has this requirement somehow already embedded in its own structure. For instance, when one argues that a person 'needs' something, or that the other person is 'vicious and depraved', not only does she entail that this is true for her, but

implicitly she means that this must be true for other people as well. In other words, it is the intrinsic meaning of words and sentences that are used in moral appraisals that calls for the presence of a general agreement on the moral judgment that is being carried out by the speaker. In this aspect there lies the universalistic trait of moral judgments.

Such a universalistic trait may be deemed as close to the notion of impartiality (see e.g. Barry (1995)). In fact, the impersonal and the impartial perspectives have often been associated, probably since Harsanyi's (1977) well-known reconstruction of the Smithian idea of the impartial observer. In this setting, the view taken by the impartial observer, i.e. he who assigns equal weight to the interests of any individual present in society, can be said to take an impersonal perspective, which leads to the utilitarian moral principle. However, the two characteristics are not necessarily equivalent. As Wiggins (1995: 61) argues, impersonality may require a *partial* treatment of alike situations, for instance by endorsing a view of morality based on of self-referential altruism, or by prescribing a more favourable to one's family, friends, or compatriots than other people. In all these cases, the impersonal view would be in opposition with an impartial one. Moreover, Wiggins starts from the definition of impartiality proposed by Hare (1952), - i.e. the principle that "equal consideration is always to be given to equal interests, whosever those interests are and whatever our relation with or nearness/distance from the person who has these interests" – in order to clarify the relationship between the two. His conclusion is that impartiality cannot be relied upon to build on that the whole sense of morality, but can at most act as a 'test' of our moral ideas *after* a sense of morality, which ultimately comes from impersonality, has already become established. In other words, impartiality cannot be constitutive of morality, but it turns out to be one of its necessary characteristics.

The second aspect of the Humean moral theory, i.e. that relating to the fact that moral judging is ultimately a *subjective* activity, is brought out clearly when Hume explains that morality is a matter of individual sentiment. In another section of the *Inquiry*, he argues that, *"The notion of morals implies some sentiment common to all mankind, which recommends the same object to general approbation, and makes every man or most men agree in the same opinion concerning it. It also implies some sentiment, so universal and comprehensive as to*

extend to all mankind, and render the actions and conduct even of the persons the most remote, an object of applause or censure, according as they agree or disagree with that rule of right which is established. These two requisites circumstances belong alone to the sentiment of humanity here insisted upon." The emphasis on sentiments, which makes for Hume's 'anti-rationalist' approach to morality, is spelled out even more clearly in the *Treatise* (1740: 468): *"The vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your own reflexion into your own breast, and find a sentiment of disapprobation, which arises in you, towards his action. Here is a matter of fact; but 'tis the object of feeling, not of reason."*

However, it would be wrong to interpret Hume as a predecessor of the *emotivists*, as could be implied by this treatment of moral sentiments. In fact, it is precisely the impersonal perspective that individuals adopt when making moral judgments attributes a universalistic content that is incompatible with emotivism (Le Caldano (1991)). Justice is thus perceived by individuals through their sentiment, not through their reasoning, so that the procedure of adopting the common point of view is not based on rationalistic investigation, but on introspection of one's own sentiments.

This analysis should clarify to what extent individual moral judgments can be relied upon to converge even within a subjectivist setting. Convergence is *not* a quasi-logical necessity of this account, but a *fact* – in principle open to empirical investigation - that can be relied upon to emerge because of the basic traits of human psychology. It is the natural characteristic of human beings to take on the common standpoint and form judgments having universal validity that causes moral ideas to become, to a large extent, shared. For an individual will be called to factor others' interests and ways of judgment into *her* own view when adopting the common standpoint, which makes it unlikely that individuals sharing a common background of information and cultural traits, the very background that comes from the fact of being all human beings, do not come to similar, if not coincident, ideas. As Hume wisely emphasised, this is already reflected in the characteristic of universality implicit in our language.

To be sure, individuals may be expected to have divergent ideas on practical issues and particular instances, but the agreement on the common principles can be said to hold. Moreover, even the existence of different and possibly conflicting conceptions of the good, such as religious doctrines that identify a particular notion of the good

and do not tolerate alternative notions, cannot *per se* be seen as evidence against the universality of moral judgments. For even these people who abide by these intolerant notions of the good, should come to realise that what they are endorsing is not the result of the adoption of a universal, common, standpoint, but is the outcome of a very particular perspective, which cannot be said to be universal and inter-subjective. Hence, these people may well come to recognise the same notion as others of the universal moral point of view when they take the impersonal perspective, but all the same attach a much higher weight to the pursuit of their own particular conception of the good. This does not undermine the view that the universal notion is shared among the people.

It is thus the inter-subjective character of the moral judgments, which is required by individuals adopting the impersonal perspective associated with the common standpoint, that brings about the convergence to a substantially similar set of moral ideas within individuals of a society. As Wiggins (1991: 62) argues, *"The content of morality may be given in propositions that are both well-grounded in consensus and fitted in with respect of their content and would-be universal application to make their appeal to consensus. The consensus in question there is one which it is natural for human beings living together in society to arrive at."* Morality will thus be an area in which there is nearly a 'universal consensus' among members of a community.

Therefore, in the model I will proceed to elaborate in the next Chapter, I will rely on the idea that individuals adopt a same public standard associated with the 'moral' point of view, though different motives to action, which could range from the selfish to any other kind, could count as well in the individual overall motivational system. The adoption of a unique common normative criterion valid amongst all subjects should be seen as a useful first-order approximation for the idea that individuals engaged in moral judgment will reach substantially homogenous moral ideas. In this way, I hope a viable model of individual behaviour will come into being.

CHAPTER 2

SOCIAL NORMS WITHIN INDIVIDUAL CHOICE

The purpose of the discussion put forward in the previous Chapter was to clarify some of the foundational issues lying behind a model of individual choice. The purpose of the present Chapter is to offer a formal account of those considerations, whose main outcome is the development of a utility function that responds to the presence of various, possibly conflicting, reasons to action.

I start by reviewing the 'state of the art' in decision theory (section 2.1), and the main empirical findings that emerge from experimental economics (section 2.2). Hence, after having motivated the focus on self-regarding and other-regarding motivations (section 2.3.1), I contrast the material and the ideal game as two different ways of assessing the interaction from the different standpoints associated with the two different norms of assessment, i.e. the self-interested and the other-regarding one (section 2.3.2 and 2.3.3). Then, the toolbox of Psychological games is illustrated, as well as its particular concept of equilibrium in 2.3.4. The comprehensive utility function is then presented. In 2.4 a number of theoretical models that fit in the general version given in 2.3 are reviewed. In particular, theories focussing on intensions-based motivations (2.4.1), social preferences (2.4.2), normative expectations and concern for social status (2.4.4) are illustrated. The theory of team thinking is also discussed in 2.4.5. 2.5 develops an alternative model of motivations that fits with the general framework, in which the other-regarding motivation is given by a notion of conditional compliance with a shared normative principle. This model will be contrasted in the next Chapters with the normative expectations theory, and will be

proposed as a way to solve some of the shortcomings that will be recognised in the latter. Section 2.6 offers a simple application of the model to the case of the non-profit enterprise.

2.1 A Review of the State of the Art in Decision Theory

2.1.1 Self-Regarding and Other-Regarding Motivations in Economic Modelling

The idea that individuals take into account a large number of reasons to action when making decisions, which extend well beyond the stereotypical self-interested motive, is now largely accepted among students of rational choice. As Binmore puts it (1994: 19), *"not even in Chicago are those views* [that homo economicus strictly abides by her own self-interest] *given credence any more"*. This set of supplementary motivations may vary and include motivations such as altruism, the willingness to act in accordance with the received sense of morality, the want to conform to the behaviour or the expectations of the other members of the community, or even less grand motivations such as, say, anti-conformist and purely whimsical ones.

Harsanyi was probably the first author in contemporary Economics who took on the issue of multiple reasons to action (Harsanyi (1969)); he introduced the distinction between *economic gain* and *social acceptance* as the pair of dominant interests explaining people's behaviour. Likewise, Bicchieri (1990: 838) stresses that *a longstanding tradition in the social sciences contrasts instrumental rationality and social norms as alternative ways of explaining action.* In particular, the relevance of the latter reason to action had been stressed with vigour in psychological and sociological investigations (for instance, Coleman (1990)). In terms of the taxonomy put forward in section 1.2, both accounts include a self-regarding reason to action, which boils down to self-interest, and an other-regarding one, which depends in some way on social norms, and may consist of the willingness to conform to the general norms reigning in a society, or to a search for *social status.*

Indeed, many other contributions within rational choice theory comprise both types of reasons to action: for instance Pettit (1990: 726) reduces the second motive to

an *indirect* form of *self-interest*, whereby the agent contemplates the *esteem, affection, or pleasure* with which other members of society view her actions, and which can be added to the *direct* form of self-interest, of more direct economic significance and in principle measurable in monetary terms. The two types of interests, the *economic* and the *social*, make up the *overall interest* of the individual. Sugden (1998a) upholds a very similar view, in which the self-interested (or material) motive is weighed up with a quest to live up to the expectations of other agents, expressed in terms of their expected material payoffs. Margolis (1990) argues that an 'optimal' balance between the two motives to action can be found by means of a properly 'economic' calculus, taking into account the material and immaterial resources that each agent can freely transfer between the self-interested and the social goal. Furthermore, a 'Darwinian' argument of selection *between* and *within* groups makes it possible to state in general terms such a principle of optimality, depicted by the maxim "neither selfish nor exploited" (Margolis (1990: 824)). Ben Ner and Putterman (1998) introduce a third motive to action, the *process-regarding* one, which adds to the self-regarding and the other-regarding ones to give a better account of the *values* a person takes into account when making decisions. For the latter term takes into account not only the outcomes that are obtained, but also the ways in which those outcomes are reached, thus including a specific 'moral' aspect into the objective function.

However, Elster (1990: 872) directly criticizes this view by arguing that when the economic and social motives are contrasted, then the difference between means and ends becomes blurred. In fact, some social norms – possibly, all of them – are actually identifiable through the very means used to reach the desired outcomes[1]. Hence, he links the rationality of behaviour to the pursuing of self-interest, since this would be the only case in which such a distinction can be neatly maintained. Other scholars put forward a similar argument, by claiming that these two motives to action are incommensurable, thus leading to the impossibility of the unity of practical reason (Copp (1997)). Finally, some authors adopt an intermediate stance, arguing that

---

[1] This argument may be opposed by considering that the same outcome reached though different means can be actually split into a set of different outcomes if the agent attributes intrinsic value to the

although a direct comparison is not always possible I may think of the social viewpoint as imposing some constraints to the individual self-interested choice (Rabin (1995)), or simply attaching an extra-value to the self-regarding payoff (Sacco (1997)).

In what follows I shall take on the view that it is generally possible to separate different, possibly conflicting motives to action in practical rationality, one referring to the individual sphere and the other to the social one. Indeed it seems that all the various dichotomies presented above can be pooled into two broad classes, according to the *standpoint* used in assessing the interests of an agent. Hence, I shall make use of the terms *self-regarding* – or *self-interested* - and *other-regarding* motive to action to represent the two categories composing the individuals' motivations, hoping to synthesize with these general terms the spirit of the contributions set out above.

2.1.2 Rationality and Multiple Motivations

The accounts illustrated in the previous section leave as ample as possible the range of an agent's possible motives to action. In other words, there is no constraint on the set of ends that the agent may wish to pursue. I now have to tackle the question of rationality of behaviour, which until now has been only skimmed over. The problem of commensurability of values, which was dealt with in section 1.3, was only preliminary to this issue.

In the modern approach to rationality the only requirements that a choice needs to satisfy in order to be considered rational are merely those of internal consistency (see for instance Hogarth and Reder (1985); Hargreaves-Heap *et al.* (1992)). In particular, when a sequence of choices made under different circumstances – namely, under different values of the 'parameters' that frame the context of choice – fulfils the basic axioms of transitivity, completeness, reflexivity, and possibly some others, then the internal consistency and thus the rationality of the agent can be said to be fulfilled. The utility function does not have any intrinsic meaning if not for acting as a formal device to represent such a coherent system of choice (see also section 1.1.1). In particular, individual rationality is not assessed on the grounds of the agent's

---

means used to reach the outcome. Of course this same argument can be applied to Ben Ner and Putterman's approach. See also section 1.2.2.B.

effectiveness in pursuing her self-interest, but rather on the logical internal coherence of her choices with respect to her ends; hence, even the behaviour of a saint can be assessed in terms of rationality in much the same way as that of homo-economicus.

Therefore, the focus in this approach is rather on the correct and comprehensive specification of the set of ends the agent is supposed to aim for: this is the stance taken for instance by experimental – or behavioural - economists, who strive to accommodate the pieces of empirical evidence found out in laboratory experiments. In their specifications, the standard self-interested motive is 'augmented' by a variety of 'social preferences' or 'intentions-based' motivations that make up an 'extended' objective function. In the former case, agents' utility function also depends in some way on the payoffs distribution amongst the group of people the agent is interacting with. This may lead to different specifications, such as aversion to inequality in surplus distribution, some form of altruism, or concern for one's own individual position within the payoffs ranking. In the latter case, agents are prompted to replicate the 'intention' perceived in others' actions, which clearly builds on Rabin's seminal model of fairness (1993). Some of these models will be reviews later on in section 2.4.

Therefore, the idea that people's practical behaviour is led by a variety of motivations, which may go against self-interest, is not inconsistent with the tenets of classical rational choice, although economists usually consider self-interest as the sole relevant motivation. However, some authors, notably Binmore (1999; 2002), criticise this approach arguing that in this way economic modelling becomes subject to arbitrariness, and that every social phenomenon could in principle be explained through the choice of an *ad hoc* objective function. Moreover, Binmore argues that self-interest-based modelling, although at variance with many particular pieces of evidence, still remains the best viable model for a *comprehensive* theory of individual behaviour. That is, multiple motivations model of choice can only accommodate limited factual evidence, and they are likely to cause large mistakes in forecasting individual behaviour when applied to different, more general, situations.

Although these seem sound criticism worth of being taken into account, in my opinion they do not seriously compromise the 'augmented' utility approach that will be proposed. As far as the first criticism is concerned, it suffices to say that enlarging

the set of motivations in individual choice does not make the whole problem of rational choice void. On the contrary, the traditional axioms of formal rational choice would still impose a significant burden to fulfil. For instance, Andreoni and Miller (2000) have tested – obtaining a positive answer - whether the behaviour of 'altruistic' individuals satisfied such axioms. Secondly, even admitting that the self-interested model fares better than the others as a general model of behaviour, which is in any case to be demonstrated, the need for particular models applicable to specific situations still makes the alternative approach worthwhile.

2.1.3 Bounded Rationality as an Alternative Model of Choice

The previous approach to 'reforming' theory of rational choice must be contrasted with that of bounded rationality. Scholars endorsing this approach argue that cognitive and/or informational limitations seriously constrain the actual choices made by individuals, so that the endeavour to construct a general model capable of spanning the whole set of human interaction is bound to fail from the very beginning (see North (1990: ch. 3); Nelson and Winter (1982)). However, even in this group the variety of approaches are, to say the least, copious. Some theorists model human behaviour as consisting of routines and rules of behaviour that are persistent, unless – possibly occasional – gathering of information reveal the availability of better ways of conduct. This approach, which comes close to Simon's (1955) seminal ideas on *satisficing rationality*, thus makes individual choice nearly rudimentary, and the focus is more on the learning process that determine a switch in the action rather than on the decisional process itself (see Anderson *et al.* (1988); and Holland (1974)).

Other theorists are instead more optimistic as to the possibility that, *in the long run*, bounded rationality behaviour 'converges' toward optimising behaviour. This has been even 'proved' in analytical terms by students of evolutionary game theory (Weibull (1995)). By showing that an Evolutionary Stable Equilibrium is nothing more than a refinement of a Nash equilibrium the much wished-for 'revolutionary' character of evolutionary approach has been somehow smothered (see Fudenberg and Levine (1999); Friedman (1998)). In this setting, the replicator dynamics seems to offer a

deterministic viable approximation, though in the long run, of many stochastic processes, thus further enriching the circumstances that these models would be able to account for (Weibull (2000)).

Finally, other scholars, who would well define themselves as endorsing a bounded rationality approach, seem to adopt an approach closer to the standard as far as the usual techniques of maximisation of an objective function are concerned. The approach is characterised by the attempt to model formally the informational and cognitive constraints typical of bounded rationality, so that their emphasis is more on the aspect of decision rather than on that of learning (see Sargent (1993); Rubinstein (1998)). The discussion of these two different approaches should have made clear that they are not, in principle, incompatible. In fact, the model that will be presented in this Chapter fits into the category of multiple motivations model, and it will be squared with a bounded rationality approach in Chapter 4.

2.2 Stylized Facts Emerging from Experimental Evidence

Although some economists still express doubts about the relevance of Experimental Economics[2], it is now undeniable that a theory of individual choice is required to come to terms with the evidence emerging from this field. In fact, most of the accounts that will be reviewed in this section can be seen as attempts to rationalize the results of some experiments. Surprisingly enough, though, the models that have been built in this fashion do not have the required generality to cover the different 'facts' that emerge in experimental economics, and often they rely on notions of 'morality' that, arguably, are quite narrow in scope. It is the purpose of the present section to offer a brief review of the available evidence, which will be used in later sections to appraise the different theoretical models and to provide support to an alternative theoretical model that I will develop in section 2.5. The present survey is based on Fehr and Schmidt (2001) and Dawes and Thaler (1988).

---

[2] In the words of Hogarth and Reder (1986: 5) *"when faced with such evidence,* [i.e. that coming from experiments] *economists are forced into 'rationalizing', discrediting, or arguing the irrelevance of the empirical findings"*. For a specific example of this kind of behaviour, see Grether and Plott (1979).

2.2.1 Heterogeneity of Individuals and the presence of non-selfish motivations

Even though it may seem unnecessary to stress, the first clear fact to emerge from experimental evidence is the large variety of behaviours from individual involved in experiments. In particular, nearly every experiment sees the coexistence of a group of subjects who appears to be typically selfish, and a group of agents whose behaviour is non-selfish. There is then no doubt that the self-interested hypothesis cannot represent a comprehensive descriptive model of human behaviour. Moreover, non-selfish behaviour presents a wide diversity of shapes, and appears to spring out of different and possibly conflicting motivations rather than from a single source. Hence, a theory that aims to offer a valid account of 'real' human behaviour should be comprehensive enough to allow for a wide range of 'types' of behaviour and of prescriptions of actions in particular situations.

2.2.2 Altruism

Another unquestionable fact is that altruism represents an important motivation for a non-negligible fraction of agents. Real life is full of examples in which this is manifest; the annual comic relief collects huge sums of money, customers give tips to waiters even in restaurants where they will never come back, people hand over to the police wallets full of money found in the streets. Hence, it does not come as a surprise that the same kind of behaviour emerges in experiments that reproduce the previous examples.

The setting in which this result can be verified is the so-called Dictator Game[3]. Here, there exists only one active player, the Allocator, who has to divide a fixed amount of money – generously provided by the experimenter - between her and a Recipient, who has no power of either changing such an allocation or possibly refusing it - as instead occurs in the Ultimatum Game. A selfish behaviour would clearly prescribe to the Allocator not to leave anything to the other agent, but experimental evidence shows that this is not the case. In every experiment conducted in this

framework there exists a non-negligible quantity of individuals who donate positive amounts of money to the other party, their percentage ranging from a minimum of 10 percent - e.g. Hoffman *et al.* (1994) – to a peak of 100 percent (Cox (2000)). In the latter case, the Allocator knows that every sum transferred to the Recipient would have been multiplied by a factor of 3 by the experimenter.

That altruism can sometimes even overcome self-interest has been shown in a slight variation of the Dictator Game tested by Charness and Rabin (2000). The Allocator has here the choice between two options. The first is perfectly egalitarian, i.e. (400,400), where the first number denotes the payoff for the Allocator and the second that for the Recipient. The second attributes a much higher benefit for the Recipient in exchange of some cost for the Allocator: (375,750). It turns out that 49 percent of the Allocators choose the latter option, thus showing their availability to sacrifice their self-interest in exchange of a substantial extra surplus for their party. Such a result of altruistic behaviour amongst Allocators seems a robust feature of the experimental literature, as similar results can be found, for instance, in Andreoni and Miller (2000). In particular, they show that 20 percent of their agents behave as 'surplus maximisers', in that they are available to give up shares of money initially attributed to them if this is compensated by a more than equivalent gain for some others.

However, that altruism is only one of several components of the multifaceted sphere of other-regarding motivations can be shown by looking at some further results that obtain upon modifying the setting of the experiments reported above. In particular, Charness and Rabin note that it suffices to increase the relative payoff attributed to the recipient with respect to that assigned to the Allocator to reverse the previous result of altruism. Thus, when confronted with the choice between the two options (400,400) and (400, 2000), 62 percent of the Allocators now opt for the first alternative. The most reasonable conclusion seems to be that a quarter of the 'supposedly' altruistic agents according to the first version of the experiment, are so only inasmuch as the relative difference in their income does not exceed a certain

---

[3] See note 1 in the first Chapter for the indication about the gender of the players. This applies to the whole of this section

threshold, beyond which a concern for egalitarianism, or a sentiment of envy for the other's income, causes such an altruistic spur to collapse. Likewise, Andreoni and Miller find that 30 percent of their agents behave as egalitarians in their experiment.

In fact, some critics put in doubt the possibility of generalising the results obtained in Dictator Games, because of their instability in the face of even slight variations in the experimental settings or procedures, and because of the rareness of situations alike Dictator Game in real-life situations, where it is more common that both parties have some degrees of power in the determination of the income distribution (Fehr and Schmidt, (2001: 30)). Nevertheless, altruism emerges in other experiments that are characterised by a more complex situation of interaction, such as the Trust Game and the Gift Exchange Game. The former is equivalent with a two-stage Dictator Game where each party alternates in the role of Allocator. The latter has the same structure, with the difference that it bears a closer correspondence with a 'real' interaction between an employer who chooses to pay a wage within a certain interval and an employee who can then perform different levels of effort. In both situations, a form of altruistic behaviour analogous to that of the standard Dictator Game is frequently observed. That is, both the first player and the second to move 'send' and 'return' positive amounts of money to each other in the course of the interaction (see e.g. Berg *et al.* (1995); Fehr *et al.* (1993)). However, the problem with such experiments that are more complex, and admittedly more akin 'realistic' situations, than the Dictator Game is that the form of altruism therein observed cannot be said not to be influenced by other types of motivations, and in particular by the willingness to reciprocate others' intentions, as will be stressed in the following sections.

The same type of problem arises in relation with the Public Good Game, where a group of individuals have to submit individually costly contributions for the provision of a public good, and the possibility of free-riding is incumbent. Reciprocity can seemingly be excluded as the relevant cause eliciting a *co-operative* behaviour, at least to the extent in which positive levels of contribution can be observed even in single trials of the game, so that the public good is produced in a proportion ranging from 40 percent to 60 percent of its optimal level (Dawes and Thaler (1988)). In this case, so the argument goes, given the absence of previous occurrences of the game, the co-

operative behaviour cannot be said to be elicited by the willingness to replicate others' behaviour. However, this argument overlooks the possibility that agents apply a form of reciprocal behaviour in relation with their *expectations* about others' behaviour. In other words, an individual disposed to reciprocate could co-operate even in single trials of the game because she *expects* others to do the same. This line of argument is consistent – but not necessarily coincident - with the idea that individuals have a disposition to apply a 'norm of co-operation' 'hard-wired' in their motivational system from real-life situations, and are *conditionally* disposed to abide by this even in the experimental context. I shall come back to this interpretation later in this section. Other kind of explanations appear possible, and discerning whether the observed behaviour is due to a purely altruistic motive or to a more complex form of group-oriented or group-regarding motivation (for a discussion, Dawes and Thaler (1988)), or even to a form of *impure* altruism (Andreoni (1989)), is difficult.

Hence, I think it is sensible to take the evidence presented in this section as supporting the idea that an altruistic trait is a relevant component of the motivational sphere of many individuals, although this motive can be offset by other concerns, such as fairness and/or envy even in the same 'idealised' situation of a Dictator Game, or by other motives such as reciprocity or group-oriented motivations in more complex interactions.

2.2.3 Concern for Fairness - and Morality

The first experimental result that attracted the attention of many economists and psychologists is certainly that obtained by Guth *et al.* (1982) in the Ultimatum Game. The structure of this game is similar to that of the Dictator Game outlined above, with the only difference that after the 'Proposer' has made an offer to the 'Recipient', the latter has now the power to accept or reject such an offer. In the first case, the two players receive what was prescribed by the Proposer's offer, whereas in the second case both players receive nothing. Applying standard Game Theoretical analysis to this interaction, thus also accepting the assumption of common knowledge of players' rationality, the predictions are rather sharp: the Receiver will accept every positive

offers made to him thus the Proposer will offer the smallest possible sum to the Receiver and keep the rest for herself.

A robust result in experimental economics across hundreds of experiments is that *both* predictions are systematically refuted. Thus, Receivers reject positive offers and Proposers assign larger than the minimum shares to Receivers. In particular, proposals offering less than 20 percent are rejected with probability 0.4 up to 0.6, and there is an evident negative correlation between the size of the offer and the probability of rejection (Fehr and Schmidt (2001: 5)). Besides, not only is it often the case that the modal offer turns out to be the 50-50 split - e.g. Guth *et al.* (1982), where this has a frequency of one third of all the offers -, but also it has been shown (Okuno-Fujiwara *et al.* (1991)) – that what turns out to be the modal offer – which is always different from the all-nothing split predicted by game theory - is the action maximising the expected income of the Proposer.

Modifications and different treatments of this experiment have tried to assess the causes of these results. This is more complicated in account of the Proposers' behaviour, as *both* fairness and self-interested reasons could support this behaviour. In fact, a significantly high offer by the Proposer could either mean that she is concerned with re-distributing part of the overall wealth to the other party, but also she may fear seeing her offer rejected. Experiments that have tried to discriminate the two components confirm that Proposers are not uniquely fair-minded. For instance, Forsythe *et al.* (1994) compare the results of a Dictator Game and an Ultimatum Game and point out that offers to the Receiver in the former are much lower, though positive, then in the latter. This implies that Proposers apply backward induction and offer some money for self-interested reasons.

In contrast, it is apparent that some non-self-interested argument must be embedded into the motivation of Responders. In general, the motive that Responders indicate as their reason for turning down positive, yet 'low', offers is that they perceived such offers to be 'unfair' (Fehr and Schmidt (2001: 6)). However, in spite of the term 'fairness' being referred to as a nearly 'intuitive' concept in most of this literature, I believe it is not altogether clear what Responders in Ultimatum Game, and more generally individuals, deem as 'fair'. So, though it is true that the perfectly

egalitarian solution is clearly a focal point and could be easily conceived as 'the fair' allocation within this context, it is also true that Responders seem to be willing to accept lower offers than this. In fact, they seem to be happy to have a non-negligible share of the pie, although this may be much less than the 50 percent. For instance, Kahneman *et al.* (1986) determine the mean minimum acceptable offer as ranging between 20 percent and 26 percent of the entire amount *m*.

But the fact that individuals may have different concepts of fairness is neither the only nor the major problem with this explanation. An interesting but not much investigated issue is whether individuals identify fairness as affecting only their *self* in relation with the others, or if they assess fairness in terms of the overall distribution amongst *all* players. In other words, the issue is whether single individuals assess fairness assuming a *personal* - their own, or possibly that of some other subject - or an *impersonal* standpoint. As we shall see, some authors like Fehr and Schmidt (1999) believe in the first option, and assume that individuals compare their own payoff with those of any other individual involved in the interaction to gauge the overall fairness of the outcome (see below). Some other authors – notably, Guth and van Damme (1998) and Ockenfels and Selten (1998) - take a somehow intermediate position by assuming that the standpoint to appraise fairness is always personal, but that individual takes as a reference point the *average* allocation for the other players. Thus, some form of impersonality emerges in the second term of the comparison. Perhaps surprisingly, though, no author has up to now embraced an *impersonal* notion of fairness as the relevant one for the individual. It is instead this latter route that I shall pursue in this Chapter.

I shall leave aside for now the theoretical considerations and analyse the limited empirical evidence on this point. A way to test whether the concept of fairness that individuals use is impersonal in character is to consider if rejections in the Ultimatum Game can be motivated by a supposedly unfair treatment towards a third party not directly involved in the action. A setting to test this is a variation of an Ultimatum Game where the Proposer is now called to decide how to share the pie between herself, a Receiver, who has as usual the possibility to accept or reject the offer, *and* a dummy player, call it the Beneficiary. Guth and van Damme (1998) have analysed this

game in an experimental setting, and report that little or nothing is generally given to the Beneficiary, whereas substantial shares are assigned to the Receiver. What is more, they state, *"there is not a single rejection that can clearly be attributed to a low share for the dummy"*. This would buttress the notion of a personal, self-based, standpoint in the assessment of the fairness of an outcome.

However, this result is completely overturned if less abrupt ways of 'punishing' the Proposer than rejecting the whole allocation is conceded to the Receiver. Fehr and Fischbacher (2000) consider a game in which a Spectator observes the allocation assigned by an Allocator to a Receiver in a Dictator Game. The Spectator is endowed with a sum of money, which she can only spend in punishing the Allocator, with a Conversion rate of 3:1, i.e. for any unit of money spent by the Spectator the monetary payoff of the Allocator is reduced by three units. A punishment would clearly support the view that individuals do not necessarily adopt a self-based notion of fairness, but put themselves in the shoes of other agents to evaluate the fairness of an allocation. The result is that punishment can indeed be often observed. For example, if the Proposer gives nothing her income is reduced by roughly 30 percent. Fehr and Fischbacher conclude that, *"this indicates that many players do care about inequities among other players"*. Oswald and Zizzo (2000) also arrive at the same conclusion that subjects care about the inequities among the set of reference agents and are available to sacrifice their own money in order to punish a behaviour viewed as unfair.

Hence, there seems to exist non-controversial evidence in favour of the idea that individuals often adopt a non-self-based standpoint in evaluating the fairness of an allocation, and punish others' behaviour even if this is individually costly in monetary terms. However, this is *per se* only conducive to an *inter-personal* notion of fairness, in that an individual adopts the perspective of another agent and assesses whether the payoff that this particular agent receives can be considered fair. To obtain a fully *impersonal* notion of fairness, not only an agent should take the standpoint of any particular agent involved in the interaction, but also she would need to adopt what Hume calls the 'common viewpoint' to all humanity (see section 1.4) and reach an *impersonal* standpoint to appraise whether the *overall* allocation can be considered to be fair.

Some evidence on the nature of the perspective taken on by individuals when judging the fairness of an allocation can be drawn from looking at Public Good Games. Here the behaviour of a defector can be deemed as going against a notion of public interest, rather than against the interests of a player in particular. Hence, a player who punished a defector in this context may be deemed as acting in favour of a clearly identifiable notion of common interest, as she cannot expect any future benefit from this action. Arguably, she thereby could be said to have taken an *impersonal* perspective, and adopted that action able to restore the common good. To be sure, this is only partially true. The action of the defector harms not only the common interest, but also the self-interest of the agent. It may well be true that the punisher of a defector thinks to be acting to defend her *own* interests rather than the common interest. More likely, the two components – the self-interested and the group-oriented - are both present and probably reinforce each other. To discriminate between the two hypotheses one would need a careful designed experiment in which the interests of the group differ from the interest of the self, but, to the best of my knowledge, this has never been attempted so far.

However, although these two components cannot be separated out so easily, the frequency and the impact of punishment is so significant in Public Good Games that the hypothesis that individuals are concerned with the *common* interest, rather or along with their *own* interest, can be thought of as receiving some support. In fact, a notable characteristic of repeated Public Good Games is that the degree of Co-operation remains positive throughout the trials of the game – usually 10 repetitions, but *declines* sharply over time (Dawes and Thaler (1988: 189)). For instance, in Isaac *et al.* (1985), the contribution rate starts out at 53 percent of the optimal quantity in the first trial, and dwindles to a bare 16 percent after only five trials. This is a rather surprising feature on which I will comment in the next section. What is important to stress now is that the possibility of punishing changes radically such a result. Fehr and Gachter (2000) introduce a 'punishment' stage after each repetition of the game, where a subject can give up a unit of her income to reduce the income of another subject by three units. This has the effect of maintaining the *average* contribution rate at the remarkably high level of 75 percent of the optimal quantity. If the subjects are allowed

to stay together for all the periods, the co-operation rate in the final period even reaches 90 percent. Carpenter (2000) also shows that with a larger group size than that of the previous experiment – ten people rather than four – subjects achieve almost full co-operation even with a random group composition over time.

Punishment is not the only way by which high rates of co-operation can be reached within a group; a less brutal way is through talk. Van de Kragt *et al.* (1983) conducted an experiment in which the contribution of a fixed number of agents is sufficient to provide the public good. However, since individual contributions are costly, the free-riding problem persists, and it is somehow made worse by the presence of a co-ordination problem. In their words, the *fear* of spending their money in vain in case an insufficient number of agents contribute adds to the component of *greed* in desiring to free ride on others. However, they show how allowing people to talk before the experiment is run permits to reach the provision of the public good in all of their trials. To be sure, the pre-talk stage acts as a powerful device to 'solve' the co-ordination problem and form strong expectations on the fact that the subjects who have been 'designated' to volunteer will in fact co-operate. In fact, since every designated individual is pivotal for the provision of the public good, she will in fact find in her interest to co-operate, provided that she holds high enough expectations that the other designated contributors will contribute, too. However, even in this case the evidence is not contrary to the supposition that agents embrace an *impersonal* notion of fairness, which possibly overlap with the self-interested motivation. Besides, as emphasised by Elster (1986), the value of discussion may lie in that it 'triggers' ethical concerns that yield a utility for doing the 'right' thing, i.e. a form of impure altruism. In another experiment van de Kragt *et al.* (1988) test these hypotheses by means of a setting in which various groups of people face a standard public good problem, i.e. where any contribution can enhance the provision of the public good, so that it is socially optimal to pay the entire endowment single individuals have. Two different treatments are put in place. The first consists of allowing discussion or not; the second concerns the possibility of doing side payments to the subjects belonging to the alternative group in which subjects are divided. The positive impact of pre-talk is confirmed even in this case: discussion rises the contribution level from 30 percent

to 70 percent. However, this only holds when subjects believe that the money goes to members of their own group; that is, if the payments of the members of a group finances the public good for the *other* good, the contribution rate shifts back to 30 percent. The conclusions van de Kragt *et al.* draw is *against* the impure altruism hypothesis. If this was true, so they argue, discussion should elicit co-operation even when directed to the members of the other group-which consists, after all, of very similar people who were indistinguishable prior to the random drawing. The fact that this does not occur at all, seems then to enforce the view that people do not follow the altruistic action in *deontological* terms, but they are interested in the particular outcome this brings about.

Likewise, promising does not seem to bound the behaviour of individuals, *unless* it is made by any member of the team. In fact, in a second series of experiments van de Kragt *et al.* find that *"in groups in which promising is not universal there was no relationship between each subject's choice to cooperate or defect and (1) whether or not a subject made a promise to cooperate, or (2) the number of other people who promised to cooperate.* Conversely, the relevance of 'group identity' is strongly confirmed by this and previous data, and accords well with previous psychological research on the 'minimal group' paradigm (e.g. Turner and Giles (1981)), which maintains that allocative decisions can be substantially altered by even marginal manipulations of the setting in which agents interact. Instead, the fact that subjects feel bounded by universal promising clearly supports the view that their behaviour has group-oriented characteristics, as clearly universal promising creates – or reflects-group identity.

2.2.4 Reciprocity or Intention-Based Motivation

That a concern for fairness cannot exhaust the whole set of non selfish behaviour reported in the previous chapter, can be easily shown by looking at a slight modification of the Ultimatum Game. In fact, it is a robust experimental result that the degree of rejection of offers decreases significantly as the proposal of how to split the pie between the two parties has been made by a computer rather than a human being -see e.g. Blount (1995). Likewise, Falk *et al.* (2000b) study an Ultimatum Game in which there are solely two alternatives for the proposal. It turns out that the

Receiver who faces the rather unequal offer (80,20) – where the first (second) denotes the payoff assigned to the Proposer (Receiver) – refuses it with higher probability when the alternative is (20,80) than when it is (50,50).

What these results show is that individuals are not only concerned with the final *outcome* of the game, but also on the perceived intention of other individuals involved in the game. An agent can then ascribe a different value to the other players' actions on the grounds of what is, in her view, their *attitude* in carrying out that action. Hence, the same allocation of (80,20) will be rejected more frequently when the alternative is a 50-50 split than when it is (20,80), because the Receiver will interpret the Proposer's action as unkind. In this respect, a widespread hypothesis is that generally referred to as *reciprocity*, according to which a key trait in human behaviour is the willingness to reciprocate the intention perceived in others' behaviour with an action of the same 'sign'. That is, human beings are disposed to exchange kind actions with kind actions and *vice versa.* This can be specified in different ways according to the context.In particular, in bilateral interactions reciprocity comes down to what has been called *reciprocal altruism* (Dawes and Thaler (1988)), and which ultimately is the application of a tit-for-tat strategy (Axelrod (1984)). In more general contexts, such as the n-player Prisoner's Dilemma, the hypothesis of *intention-based reciprocity* assumes that an individual is motivated to act against or in favour of an agent depending on whether her action is 'fair' with respect to the interests of the whole group of agents. The assumption is that, were the intention of an agent perceived to be fair (unfair), then one relevant motivation for another agent would be to reward (punish) that action. Intention-based reciprocity can thus be seen as a more general assumption than simple two-by-two reciprocity.

Experiments have been carried out to test *either* hypotheses of 'positive' and 'negative' reciprocity. The evidence seems to support both, although some controversial results have been obtained for the positive reciprocity hypothesis. In particular, the frequent observation of punishment that has been reported in the previous section may be evidence of a willingness to reciprocate unfair actions. What experimenters have tried to test is whether this is due to a mere *taste for punishment*, i.e. a willingness *to hurt* a player who have behaved unfairly, or an instance of *inequality*

aversion, i.e. a willingness *to restore* a more equitable allocation of the payoffs amongst the agents through the punishment. The evidence seems to favour the *first* hypothesis, although the second plays a relevant role as well.

This has been proved by Falk *et al.* (2000a) in a Public Good Game in which the conversion rate of the punishment is only 1 to 1; that is, a player willing to punish has to give up one unit of her income to reduce of one unit another subject's payoff. A pure concern for the fairness of the allocation would then imply that no punishment is carried out in this situation, as such an operation would leave the relative distance between players, and thus the overall fairness of the allocation, unchanged[4]. But that this is not the case is proven by the fact that 25 percent of the subjects do in fact punish free riding even under these conditions. Conversely, since the percentage of agents punishing free-riding only rises to 36 percent when the conversion rate of the punishment is, as above, 3:1, they can conclude that nearly 70 percent of the punishing behaviour is motivated by the desire to harm the disloyal agent rather than because of inequity aversion.

The evidence seems instead more controversial with respect to the *positive* reciprocity hypothesis. The fact that in several Trust Games and Gift Exchange Games – e.g. Berg *et al.* (1995) for the former and Fehr *et al.* (1993) for the latter game - one observes a high degree of correlation between the size of 'nice' responses by the Receiver and that of 'nice' offers by the Proposer seems direct evidence in favour of the reciprocity hypothesis. However, since the same behaviour could be rationalized in terms of inequity aversion of the players (Fehr and Schmidt (2001: 34)), one needs to contrast directly the two hypotheses in order to gauge their relative importance. The findings of some experimenters seem in fact to doubt the relevance of the willingness to reciprocate nice actions. For instance, Bolton *et al.* (1998) analyse a Trust Game and find that the rewards of the Receiver in favour of the Proposer are even bigger when this has only a single offer available rather than when she has many and can make nice

---

[4] This is true, at least, if one assumes that an agent views fairness as the relative income separating her from any other agent of the group (Fehr and Fischbacher (1999)), or if one compares her income with the average one and identifies the *fair* allocation with the egalitarian outcome, as in Bolton and Ockenfels (2000). The same implication may not be reached assuming, for instance, that agents were concerned with the *dispersion* of the income around the average in addition to the average itself.

offers. Such behaviour is clearly contrary to intention-based reciprocity. Other studies – e.g. Cox (2000) and Charness and Rabin (2000) – confirm the scarce relevance of positive reciprocity.

These results are somehow surprising, and in fact have been completely reversed by more recent evidence. Falk *et al.* (2000a) replicate a Trust Game analogous to that reported above, and find an opposite result; that is, nice responses are larger in size when the offer has been made by a human rather than a computer. McCabe *et al.* (2000) also arrive at the same conclusion. It thus seems sensible to conclude that even positive reciprocity does indeed play a relevant role in individual motivations, though perhaps less big than negative reciprocity.

Another fact worth of attention, which apparently has not received the deserved attention by theorists trying to accommodate experimental evidence into models of human behaviour, is the declining degree of co-operation that one systematically observes in Public Good Games. As already noticed, in single repetitions of the game, the public good is provided at about 40-60 percent of the optimal quantity, a result that appears robust to many possible different specifications (Marwell and Ames (1981)). This is *per se* a surprising result, which clearly underscores the presence of some altruistic or group-oriented traits in many individuals. Were individuals' preferences fixed or not depending on others' actions, one could expect that the rate of co-operation remain stable if not increase across repetitions of the game. However, as stressed above, this is not in fact the case, as co-operation rates decrease over the repetitions of the game. That this is not due to some effect of learning *about the experimental context* has been proved by Andreoni (1987), who finds the same aggregate behaviour once the same group of subjects repeats a second time the experiment. Another possible explanation points at repeated game effects. As made clear by Kreps *et al.* (1982), an even minimal probability that subjects face an 'irrational' individual who plays tit-for-tat even in a finitely repeated Prisoner's Dilemma, makes co-operation at the early stages of the game profitable for a self-interested individual. For by co-operating such a 'rational' player induces the other subject to co-operate, and finally defects in the last stages of the game. However, if all individuals were 'rational' in the Kreps *et al.* sense, then one should observe a zero-degree of co-operation in the

last stage of the game. Conversely, since this is observed to remains positive, such an account cannot offer a comprehensive explanation of the evidence.

The alternative hypothesis that a considerable part of the subjects are *conditional co-operators* has instead received a strong empirical support (see e.g. Croson (1999); Fehr, Fischbacher, and Gachter (1999); Offerman *et al.* (1999)). This hypothesis implies that a subject's co-operative behaviour comes to an end when she observes that others' behaviour is selfish. Hence, the simultaneous presence of selfish individuals and conditional co-operators causes the degree of co-operation to decrease over time; in fact, if *all* individuals were conditional co-operators, then subjects who did not start co-operating should switch to co-operation after observing a substantial amount of co-operation in the first stages of the game.

This highlights the relevance of another kind of 'reciprocity' than that presented above as 'negative' and 'positive' intention-based reciprocity, which seemingly comes down to punishing and rewarding unfair and kind behaviour respectively. Such an account emphasizes the possibility that individuals 'reciprocate' with respect to the idea of the common interest of the group, so that they cease to co-operate when they observe that such interests are not being pursued in the way they consider minimally satisfactory. What such a view requires is a sense of group identity, and an impersonal notion of the common interest predicated by Hume, which individuals are available to endorse to the extent that they expect/observe a like behaviour by other agents. Some scholars talk about a 'norm of co-operation' as a concurrent hypothesis to explain behaviour observed in repeated Public Good Games (Dawes and Thaler (1988: 191)). However, I believe that the implied behaviour is more general than what seems to be implied by this term, and can extend to different situations than public good provision problems. I believe that the best way to interpret this result is by saying that individuals are *conditionally* committed to abide by the *moral* prescriptions, which can be derived by adopting an impersonal point of view. Although this hypothesis has not been directly tested in experiments, the fact that indirectly it receives some support, though admittedly in a sketchy and rather fragmented way, from available evidence, makes it worth investigating from a theoretical and empirical point of view, which will be done in section 2.5.

2.2.5 Conclusions

The purpose of the present section has been to review the available evidence produced in experimental economics, and to put forward some theoretical ideas to account for them. In the next section I shall analyse how such theoretical notions can be shaped into formal models of individual motivation. I now wish to summarise the main stylised facts that emerge from the previous survey, and stress some requirements that a model of motivation should respond to.

First, individuals are heterogeneous, so that within the same group different 'types' of individuals coexist. A good model of motivations should then allow for different possible 'types' being consistent with the general model. Second, individuals seem to take into account a *variety* of motivations when making decisions. Hence, as emphasised by the behaviour of Responders in ultimatum game, they perform *trade-off* between different motivations. Thus, the average minimal degree of rejection is neither the perfectly fair outcome, i.e. the 50-50 split, nor the perfectly self-interested one, i.e. the minimum amount, but it is what seems a rough average between the two (see section 2.2.3). Hence, the resulting action may well be a 'compromise' between two different – and possibly conflicting – prescriptions, but it may also imply a 'switch' from a kind of behaviour to on completely different as a response to other agents' actions, or as a consequence of a change in the environment. The presence of such a threshold effect will emerge clearly in section 2.6 in the study of the non-profit firm case. As an example, the observed degree of co-operation in a repeated Prisoner's Dilemma affects the propensity to co-operate in future interactions. The implication of this piece of evidence for the modelling of individual motivations is that models only relying on a sole motivational source cannot offer a comprehensive account of human behaviour. Hence, since *both* reciprocity –in the positive and negative sense – and fairness seem to be relevant – but not exhaustive - components of human behaviour, a model of individual motivation should aim to factor either elements into its specification.

Third, during the review of the evidence, I have sought to emphasise how a relevant component in the other-regarding motivation could be given by the disposition to abide by an *impersonal*, or *moral*, idea, rather than an idea of fairness that

involves a comparison between the relative position of the *self* and that of other people, considered either singularly or as a whole (see section 2.2.3). Although this hypothesis has not been directly tested, it seems consistent with all the stylised facts reported above. In particular, the altruistic behaviour observed in Dictator Games and Public Good Games may be interpreted as a disposition to bring about the *moral* allocation, as perceived by the agent in the particular context she is acting[5].

The same could be said for the many instances in which some forms of inequity aversion is observed, and, under opportune conditions, for punishment as well. The well-known finding that, when given the opportunity, people manifest a strong sense of attachment to the group they are involved, so that their behaviour turns out be *group-oriented*, cannot but buttress this idea. Admittedly, no experiment has been conducted to test this hypothesis, and the indeterminacy of what the 'moral' prescription is in many cases is clearly an obstacle to empirical investigation. However, these two aspects - i.e. whether agents adopt an impersonal perspective when evaluating the allocation and its practical content - can in principle be separated and investigated each in turn from the experimental point of view. Therefore, after having developed a general enough theoretical framework to rationalise the different theoretical ideas in the next section, I shall explore the existing theoretical accounts in 2.4, and put forward an alternative model in section 2.5. This model will be based on the idea of *reciprocity* with respect to a *moral* notion, and it seems capable of offering a comprehensive account of the facts stressed in the present section.

---

[5] An interesting approach on how subjects approach this situation is that offered by Brock (1979) when he draws the distinction between *manna-from-heaven* and *non-manna-from-heaven* type of situation; in the former individuals acquire some positive level of wealth without having *deserved* it, so that a type of redistribution based on the *need* principle becomes intuitively appealing. On the contrary, in the latter situation a distribution based on the *contribution* principle appears likely to be upheld by individuals. Since the experimental context is almost by definition a manna-from-heaven situation, then this may explain why individuals have a higher inclination to give in situations like the Dictator Game, and have a keen attitude to punish in contexts where an individual *undeservedly* earns his payoff, such as free-riding in a Public Good Games.

2.3 A General Account of a Model of Choice Based on Multiple Motivations

In what follows I will offer a formal framework to deal with some of the issues highlighted in the previous section from a theoretical point of view. In particular, my goal is to elaborate a model capable of accounting for the stylised facts emerged in the previous section, i.e. (a) the presence of different types of agents, and of (b) various reasons to action, and (c) the relevance of the reciprocity element. I will also allow for (d) that moral considerations have an impact in shaping other-regarding motivations, depending on how one specifies the normative criterion of assessment of the states of affairs regarding others. As for (b), the kind of reasons to action I focus on are, as customary in this literature, self-regarding and other-regarding motivations. The first is identified as the self-interest of the agent. The second is more complex in that it involves the assessment of a social situation from an external standpoint, which may embody a moral principle, an ideological standard, a set of precepts derived from some codes of behaviour, etc. I shall generically call this a *social normative principle*, or, more concisely, a *normative principle*, or even a *normative criterion*, implying that it offers an appraisal of the *social* outcome based on some standard of assessment (see section 1.2); that is, it takes into account the consequences for each agent involved in the interaction from some normative standpoint[6].

This is the way in which social norms 'enter' the individual system of motivations; norms are here seen primarily as criteria of evaluation of a state of affair expressed from a standpoint that is different from that of the self-interest. According to the particular specification of this aspect, a norm could acquire motivational strength *in itself*, that is, because its prescriptions are compelling for a moral or ideological point of view; or the individual may see in it a way to solve co-ordination problems, or to implement the public good in case general co-operation of all of the individuals is required. Though all of these specifications can be factored into the model, in my view, they are somehow subordinated to the main characteristic of a norm as a general

principle of assessment of a state of affairs, which is carried out from a standpoint somehow more general than that of the self. Accordingly, the social normative principle offers an *ordering* of the social outcomes of which each agent is aware when making decisions, which can shape individual motivations in many different forms. Some of these will be reviewed in 2.4, where such a normative principle will not be specified if not for its general formal characteristics. In Chapter 4 it will instead be given the particular specifications of a Nash social welfare function.

In terms of the formal specification of the utility function, this will be made up of two components that fit with the two types of motivations that an agent has. The first source of utility is associated with the fulfilment of one's own self-interest, whereas the second is utility correspond to the compliance with the social normative principle. I shall call them *self-interest based source of utility* and *other regarding source of utility*, or, for brevity, *self-interested* and *other regarding utility*. This pair of sources of utility makes up the *comprehensive* utility function, and, in absence of any contrary reason, I will suppose that the two components enter the function additively.

In the following sections, utility is defined in relation with a situation of interaction amongst *n* agents, as represented in a *game*, where the two reasons to action correspond to two different analysis of the same game. The route I will take is to view the payoffs of the game as representing the *self-interested* utility of the agents. At the same time, though, a normative principle of assessment is defined over the payoffs of all the agents involved in the interaction, possibly including those of some agents that are affected by the interaction but cannot influence its outcome, i.e. dummy players. This offers a ranking of the social states on the grounds of the greater or lesser correspondence to the normative principle, so that each agent is aware of how much her action is being coherent with the social normative principle and contributes to the fulfilment of its prescriptions. Finally, those two different standard of assessment are weighed up in the comprehensive utility function, and the 'best' action is determined

---

[6] In fact, the adjective "social" is meant to differentiate a normative principle that refers only to the self to one that takes into account all of the individuals. The illustration of norm of assessment given in section 2.1 was general enough to include both categories.

in terms of the greater fulfilment of the two sources of value overall considered by the agent.

In order to distinguish the two perspectives with which the agent evaluates the game, I will call *material* game that associated with the self-interested perspective and *ideal* game the same interaction appraised by the standpoint offered by the normative principle. Accordingly, I will sometimes even refer to self-interested and other-regarding utility with the terms *material* and *ideal* utility respectively: I will consider the two pairs of expressions as perfectly equivalent. In fact, one can ultimately see the two perspectives as associated with two different types of solution of the same game, one boiling down to the usual Nash solution when only self-interest is taken into account, and the other consisting of the solution to the game if it was played co-operatively – namely, constraining individual behaviour not to necessarily perform the best action in terms of self-interest - and 'solved' in accordance with the ranking given by the social normative principle.

2.3.1 The Material Game

It is given a game G, made up as usual by a triplet of elements: a set I of players, a set of strategies S_i and a utility function U_i for each agent. Formally, $G = \{I, S, U\}$, where $S = \underset{i \in I}{\times} S_i$ defines the set of feasible strategies profiles, and likewise U is the set of vectors of utilities. Allowing for the use of mixed strategies by the agents, we can further introduce the operator $\Delta(X)$ to express the randomisations over a set of elements X. We can thus define the set of possible randomisations over the strategy sets of the agents: $\Sigma_i := \Delta(S_i)$; finally, we can consider the vector including a randomisation for each agent: $\Sigma := \underset{i \in I}{\times} \Sigma_i$, where the generic element is indicated with $\sigma \in \Sigma$.

In the game G, the utility functions represent a measure of the self-interest of the agents, thus reflecting the first type of motivations. This is the reason why we will call the related source of utility *self-interested.* They are defined, as customary, firstly over the outcomes of the games; that is to say, they are functions of the profiles of pure strategies: $\overline{U}_i(S)$. Furthermore, taking on standard assumptions regarding expected

utility, we introduce Von Neumann-Morgestern utility functions defined over mixed strategies profiles, $U_i(\Sigma)$, where

$$U_i(\sigma) := \sum_{s \in S} P_\sigma(s) \overline{U}_i(s) \tag{2.1}$$

$P_\sigma(s)$ represents the probability that the pure strategy profile s is played according to the mixed strategy profile σ. Provided that the nature of this game does not differ from standard game theoretical analysis, the relevant concept of solution would be the Nash's one.

2.3.2 The Ideal Game

The ideal game differs from the previous one in that agents evaluate the social situation from a different standpoint than the self-interested one, possibly including the evaluation of the material payoffs of other agents who are *affected* by their actions but *cannot affect* the final outcome. Hence, we introduce an *ideal* game G^* as an extension of the material game G, in which the set of players is possibly larger than in the material game thus modifying the corresponding set of utilities. Formally, this game is defined by the triplet: $G^* = \{I^*, S, U^*\}$, with $I \subseteq I^*$ and $U^* = \underset{i \in I^*}{\times} U_i$. Notice that the set of actions S is left unaltered with respect to the material game: by definition the players now included in the game are *dummy* players in the original one.

Resting upon this construction, we can now introduce the notion of the normative criterion used to appraise the social outcomes. This expresses the ranking of the social outcomes made on the grounds of the ideology, or the moral principle, that is being taken as relevant by the agents. In other words, we are assuming that it is possible to measure on some scale the *correspondence* of the social states to an ideal norm of assessment, which is represented by a function of the social outcomes. This is analogous to an *individualistic* social normative function in that it is dependent on the self-interested utilities of the agents involved in the interaction:

$$\overline{T} := \underset{i \in I^*}{\times} \overline{U}_i(S) \rightarrow R \tag{2.2}$$

Therefore, such a normative principle permits the creation of an ordering over the possible outcomes, which represents the assessment that an impartial spectator would

give to the different social situations on the basis of the relevant normative criterion. A higher value of the function T implies that the associated social outcome satisfies to a higher degree the normative criterion.

Of course, taking the structure of the game as granted, it is possible to make the function directly dependent on the pure strategy profile set S, and, also, on the mixed strategies of the game:

$$T(\sigma) := \sum_{s \in S} P_{\sigma}(s) \overline{T}\left[\overline{U}(s)\right] \tag{2.3}$$

In analogy with the individual expected utility, the expected normative function is simply a weighed sum of the welfare levels under all possible pure strategies profiles, with weights given by the probabilities that each outcome is actually played.

2.3.3 Beliefs, Individual Utility and Equilibrium

The analytical apparatus illustrated in the previous sections is common to most of standard game theory. However, for many of the applications that will follow we need an extension of this toolbox, which draws on the approach of Psychological Games (Geneakoplos, Pearce and Stacchetti (1989); GPS from now on). In fact, a key aspect of the way social norms affect individual behaviour is through the role of mutual expectations between members of a community. Typically, an individual will perceive the existence of a social norm through the net of expectations that other members of a community have on her following a particular behaviour. In other words, the individual will attach importance to the extent to which her action conforms to the *expectations* of other members of a community, which in turn are based on the existing social norm. This simple consideration highlights that what may be important, in this context, for an individual is the *difference* between what is expected from her and her actual behaviour. For instance, in some parts of Southern Italy there is still the habit for the husband not to cook or even participating in laying the table with his wife: these are in fact considered typically feminine activities, and a man who helped his wife would and attract the mockery of bystanders and probably he would be laughed at. Instead, in most other communities, helping a wife is considered a normal activity for a man to do, and it would not attract any scorn nor, because of its diffusion, any sense of approval. Thus, the same action, i.e. not cooking and laying the table, will

attract completely different evaluations depending on the expectations that are associated with it.

If the game theoretic toolbox remained that outlined in the previous sections, there would be no chance to capture this further aspect, as individual utility would only depend on *actions*. Instead, if we aim to take into account the impact of expectations and social norms on individual behaviour, utility must be sensitive to expectations as well: only in this way can the individual, say, attach a disutility to the fact of having failed to conform to a social norm, or assign some extra-utility to the fact of having gained the commendation of some other agents for a 'good' action. Psychological game theory provides us with the most general approach to this sort of issues, by developing a framework in which not only does individual utility depend on agents' actions, but also it is affected by the state of individual expectations on each other's actions.

In order to make this idea tractable from the analytical point of view, we first need to formalise the notion of belief. However, this is not an entirely easy task, as typically beliefs can reach different, possibly unbounded, *orders*. In fact, the requirement of common knowledge, which is ordinary in most economic analysis, does rely on the coherence of infinite order expectations (see Myerson (1991): Ch. 1 for some paradoxes implicit in common knowledge). We shall expand on this point further later on (see section 3.1.2.A). However, the applications that will be presented in this thesis will only require the first two orders of expectations.

Let me introduce the formal treatment of this concept, which is based on GPS (1989), and Mertens and Zamir (1985). A *first order belief* for player i is a probability measure over the other players' mixed strategy set, namely $B_i^1 := \Delta(\Sigma_{-i})$; thus the generic element $b_i^1 \in B_i^1$ indicates the probability with which i believes that the other players are going to implement the profile of strategies σ_{-i}. In the same fashion we can define $B_{-i}^1 := \times_{j \neq i}(B_j)$. Obviously, when there are just two active players, we have $B_i^1 := \Delta(\Sigma_j)$ and $B_{-i}^1 := B_j$.

A *second order belief* for an i-player is a conjecture over the *beliefs* held by other players on each other's strategies. Therefore, it consists of a probability measure over the

Cartesian of other players' beliefs of first order: $B_i^2 := \Delta\left(B_{-i}^1\right)$. In the case of only two persons being involved in the interaction, the generic element of this set, $b_i^2 \in B_i^2$, represents *i*'s probability distribution that the belief of *j* over *i*'s strategies is b_j^1[7]. Iteratively, one can define *k*-order beliefs as follows: $B_i^k := \Delta\left(B_{-i}^{k-1}\right)$. Notice that from the formal point of view, the order of beliefs *k* is not bounded, thus making it possible to deal with beliefs of infinite order on each other's beliefs. I indicate with $b_i = \left(b_i^1, b_i^2, ...\right)$ the infinite-dimension vector collecting the beliefs of each order for player *i*. Consequently, we can define $b = (b_1, \ldots b_n)$ as the profile of beliefs for each of the *n* players taking part in the game.

Drawing on this analytical apparatus, we can now give the definition of the *comprehensive* utility function. Generally speaking, this will be characterised by beliefs *entering* explicitly the arguments of the utility function. Not only will the subjective probabilities that make up (2.1), (2.2) and (2.3) now depend on the individual beliefs on other players' action, but also any other belief included in the vector *b* defined above, can affect individual utility. For instance, coming back to the above example, if the husband is aware that bystanders attaches probability p_i to him not cooking, and if he experiences some psychological cost to be laughed at when breaching the social norm prescribing him to do so, then his utility may look like:

$$V_i\left(\sigma_i, b_i^1(p_i)\right) = \sigma_i - \lambda_i(1 - p_i) \qquad (2.4)$$

Here σ_i is the probability of the husband laying the table and cooking, and the comprehensive utility represents preferences such that that the husband would like to lay the table and cook, but also dislikes to be laughed at when doing so. λ_i measures the psychological cost of being mocked, compared with the psychological satisfaction

---

[7] Although beliefs are probability distributions iteratively defined over probability distributions, the associated *marginal* probabilities over a generic opponent's strategies can be easily obtained by means of the following formulas:

$$P_{b_i^1}(s_j) = \int_{\Sigma_j} P_\sigma(s_j) P_{b_i^1}(\sigma_j) d\sigma_j \; ; \; P_{b_i^2}(s_i) = \int_{B_j^1} P_{b_j^1}(s_i) P_{b_i^2}(b_j^1) db_j^1 \, .$$

Thus the first formula indicates the overall probability that player *j* is going to play s_j, according to the belief b_i^1 held by player *i*, and the second the overall probability that player *j* holds about *i*'s performing s_i, according to the second order belief b_i^2.

of cooking. In fact, when p_i is zero, then there is the expectation from the 'community' of bystanders on the husband *not* laying the table, thus the psychological cost of breaching the 'social norm' of not setting the table will be the highest. If instead there is no such a convention, i.e. p_i is equal to 1, then no psychological cost will be attached to setting the table.

In more general terms, we shall see the comprehensive utility function as made up by the sum of two components, which corresponds with the two sources of utility illustrated so far, i.e. a *self-interested*, or *material* utility, and an *other regarding* utility. The former is associated with the payoffs in the material game (see 2.3.1) where the latter reflects the reason to action given by the conformity to the normative principles T (see 2.3.2). The comprehensive utility function will then have the following form:

$$V_i(\sigma;b)=U_i(\sigma,b)+\lambda_i g_i[T(\sigma,b)] \quad i\in I^* \qquad (2.5)$$

The vector b of beliefs enters both components of utility, as subjective probabilities on others' behaviour will coincide with first order beliefs in the vector b. While the material, or self-interested, utility U_i is 'standard' in that it is shaped in accordance with the agent's payoffs given in the material game, other-regarding utility is expressed as some function g of the social normative criterion T. g can in principle be thought of as differing amongst individuals. However, relying on the argument put forward in 1.4, it will be generally assumed to be shared by all agents. For simplicity, the two components enter the function additively, and the parameters λ_i, possibly differing across the set of agents, measure the weight attributed to their other regarding utility in the face of the self-interested source of utility. The function g may be specified in different ways in order to account for various possible forms of the other regarding motive to action.

The peculiar innovation introduced in the comprehensive utility function, that is, the inclusion of beliefs in the arguments of the function, calls for an extension of the standard concept of solution of games, namely the Nash equilibrium. We shall adopt the original notion of Nash psychological equilibrium put forward by Geanakoplos *et al.* in their seminal contribution, although some refinements of this notion have been suggested (Van Kolpin (1992)) and others will be presented in Chapter 4.

The main idea of this concept is that, if we are in equilibrium, then the beliefs of rational players must conform to the strategies that are being played. As an example, if in equilibrium I observe my opponent playing the (possibly mixed) strategy $\sigma_j \in \Sigma_j$, then my first order belief must assign probability one to that particular strategy and 0 to all of the others. This is tantamount to saying that once an equilibrium has been reached, all of the first order beliefs must be single-point distributions assigning probability one to the equilibrium strategy. The higher order beliefs are then generated upon a condition of *coherence* with this initial condition (GPS (1989: 64)). We shall call $\beta_i(\sigma)$ the distribution of beliefs associated with the distribution that is coherent with assigning probability 1 to the strategy σ, and with $\beta(\sigma) = (\beta_1(\sigma),...\beta_n(\sigma)) \in B$ the profile of such beliefs for the *n* players.

Recalling the definition of b_i as the vector collecting the beliefs of each order for player *i*, we are now able to provide the definition of Psychological Nash equilibrium (GPS (1989: 65)):

A psychological Nash equilibrium for a n-person normal form psychological game G is a pair $(\hat{b}, \hat{\sigma}) \in B \times \Sigma$ such that:

i) $\hat{b} = \beta(\hat{\sigma})$

ii) for each $i \in I$ and $\sigma_i \in \Sigma_i, V_i(\hat{b}_i, (\sigma_i, \hat{\sigma}_{-i})) \leq V_i(\hat{b}_i, \hat{\sigma})$ (2.6)

Condition (ii) is a simple restatement of the standard Nash equilibrium condition, affirming that for each player the equilibrium strategy must confer a payoff not smaller than what attained by any other feasible strategy, *given* the opponents' strategies *and* the beliefs. Condition (i) restrains the beliefs to reflect and be coherent with the equilibrium strategy. Notice that if beliefs are not part of the utility function then condition (i) becomes redundant and the definition will boil down to the standard Nash equilibrium definition.

2.4 Different Specifications of Other-Regarding Utility

Several of the most recent contributions in the field of multiple motivations utility function share the general format given by equation (2.5) in that an other-regarding

source of utility adds to a self-interested one. This is also consistent with the general specification derived by Segal and Sobel (1999) following an 'axiomatic' approach. However, significant differences exist as to what it is deemed as the relevant other-regarding motivation. In fact, most of the recent studies carried out in experimental economics are oriented to 'fine-tuning' the theoretical specifications of the other-regarding motivation with the experimental results. In the present section, I shall review some of these theoretical accounts, and highlight the extent to which they accord with such evidence.

2.4.1 Intentions-Based Motivations

This thread of literature is based on Rabin's seminal model of fairness (1993). The main idea is simple, and seeks to give a rationalisation of the evidence presented in 2.2.4. It would be a natural trait of human motivations to reciprocate the *attitude* perceived in other individuals' actions towards the self, so that an individual would be likely to respond to actions perceived as kind with a kind action, and *vice versa*[8]. On this view, Rabin's model is a formal device to incorporate these observations into individual choice theory. The theory of Psychological Games provides with some tools to embody these considerations into a formal analysis. In fact, it introduces *beliefs,* of every possible order, on each other behaviour into the utility function (GPS (1989)). In this fashion, as we saw earlier, it is possible to model the idea that an agent can be more or less satisfied depending on how others' actual action correspond to her initial expectations. In particular, for simplicity restricting the attention to the case of two-person interactions, Rabin considers a pair of 'kindness functions', which measure the extent to which the agent's own and her counterpart's actions increase or diminish one another expected payoff, on the grounds of the first and second order expectations formed by the agent. This estimate is used by each agent to appraise the kindness of the other party to herself, using in particular the second order expectation, and the kindness of the subject herself toward the other agent, as perceived by the other agent drawing on the first order expectation of what the other agent expects from herself.

The way in which these functions are constructed is to consider the best and the worst payoff that the each agent can cause to the other on the basis of the reciprocal expectations, and then to consider how the payoff actually brought about lies between those two extremes.

In formal terms, the kindness function of subject *i* with respect to her opponent is given by:

$$f_i\left(\sigma_i, b_i^1\right) = \frac{U_j\left(\sigma_i, b_i^1\right) - U_j{}^{FAIR}\left(b_i^1\right)}{U_j{}^{MAX}\left(b_i^1\right) - U_j{}^{MIN}\left(b_i^1\right)} \tag{2.7}$$

where $U_j{}^{MAX}\left(b_i^1\right) = \arg\max_{\Sigma_i} U_j\left(\sigma_i, b_i^1\right)$ and $U_j{}^{MIN}\left(b_i^1\right) = \arg\min_{\Sigma_i} U_j\left(\sigma_i, b_i^1\right)$. In other words, these are the maximum and the minimum payoffs for player *j*, which can be possibly induced by player *i*'s choice, on the basis of *i*'s first order belief about *j*'s playing, which, in accordance with the notation put forward in section 2.3.3, is denoted by b_i^1. $U_j{}^{FAIR}\left(b_i^1\right)$ is instead a supposedly 'fair' level in payoffs attribution, which Rabin models as the *middle point* between the highest and the lowest of *j*'s payoffs within the set of Pareto-efficient outcomes. The last proviso causes the highest and the lowest payoffs used to compute the fair payoff to possibly differ from those entering (2.7). Its general interpretation is thereby that the higher (the lower) the *actual* payoff achieved by *j* in relation with the fair level, the kinder (more hostile) is the perception of *i*'s action by *j*. In particular, an action will be considered as 'kind' by *j*, in *i*'s view, if and only if the kindness function $f_i\left(\sigma_i, b_i^1\right)$ reaches a positive level, i.e. *i*'s action brings about a payoff higher than the fair one to *j*, and *vice versa*. The division by $U_j{}^{MAX}\left(b_i^1\right) - U_j{}^{MIN}\left(b_i^1\right)$ ensures that such a function is an index varying between –1 and ½.

Stepping up the expectation ladder of one level, we can derive the (esteemed) kindness of *j* to *i*, in relation with *i*'s *second* order expectations:

$$\tilde{f}_j\left(b_i^2, b_i^1\right) = \frac{U_i\left(b_i^2, b_i^1\right) - U_i{}^{FAIR}\left(b_i^2\right)}{U_i{}^{MAX}\left(b_i^2\right) - U_i{}^{MIN}\left(b_i^2\right)} \tag{2.8}$$

---

[8] The evidence on which Rabin (1993: 1281) bases his account is a little more sophisticated than this: it is argued that people who are *more* inclined to reciprocate nice actions with nice actions are *at the*

where $U_j^{MAX}\left(b_i^2\right)=\arg\max_{\Sigma_j} U_i\left(b_i^2,b_i^1\right)$ and $U_i^{MIN}\left(b_i^2\right)=\arg\min_{\Sigma_j} U_i\left(b_i^2,b_i^1\right)$. *i* now puts herself in *j*'s shoes and tries to evaluate how kind *j* is being to her. Thus, on the basis of her second order beliefs, which is what *i* expects that *j* expects on *i*'s actions, *i* computes the fair payoff that she can expect, and assesses *j*'s kindness with respect to the extent in which her (expected) actual payoff exceeds or stays below the fair level.

The two kindness functions determine the following specification for the other-regarding utility:

$$g\left(\sigma_i,b_i\right)\equiv\left\{1+f_i\left(\sigma_i,b_i^1\right)\right\}\tilde{f}_j\left(b_i^2,b_i^1\right) \qquad (2.9)$$

The implication is that if the counterpart's behaviour is perceived as kind, i.e. $\tilde{f}_j\left(b_i^2,b_i^1\right)$ is positive, then agent *i* will be motivated to be nice as well, thus increasing other-regarding utility. Conversely, if the counterpart's behaviour is deemed as unkind, i.e. $\tilde{f}_j\left(b_i^2,b_i^1\right)$ is negative, then agent *i* will be willing to be as unkind as possible, so that the other-regarding term is brought to zero. (2.9) is then added to a self-regarding utility function in conformity to (2.5):

$$V_i\left(\sigma_i,b_i\right)\equiv U_i\left(\sigma_i,\sigma_j\right)+\left\{1+f_i\left(\sigma_i,b_i^1\right)\right\}\tilde{f}_j\left(b_i^2,b_i^1\right) \qquad (2.10)$$

Given the nature of 'index' of the other-regarding terms, the higher the monetary payoffs associated with the self-regarding function, the lower is the incentive to follow other-regarding considerations by the agent. Although the evidence on this point does not seem to give much support to this idea, as most experiments only find a weak effect of the size of monetary payoffs on the degree to which individuals execute non-selfish action – e.g. Cameron (1999)[9], Fehr and Tougareva (1995), this is something that we would obviously expect, and that – given the inevitably limited budget available to experiments - cannot be thoroughly tested (Thaler (1988)).

After the pioneering work of Rabin, some authors have proposed a generalisation of his model to N-agent sequential games (Dufwenberg and Kirchsteiger (1998)). The main idea is that the 'sequential reciprocity equilibrium' needs to be a fairness

---

same time those more likely to respond to unkind actions with unkind actions.

[9] This study conducted in Indonesia reports that Responders can renounce to stakes as high as three months of their salary to 'punish' a relatively unequal offer from the Proposer.

equilibrium in every subgame. It can thus be shown that *conditional* co-operation can be a sequential equilibrium in repeated prisoner's Dilemma. However, equilibria are generally multiple, and this can represent a shortcoming for the predictive power of the theory. What is more, some 'self-fulfilling' equilibria can arise that are based on apparently implausible initial beliefs. In particular, equilibria in which the Responder rejects every offer with probability one in an Ultimatum Game can be an equilibrium if the initial state of beliefs is such that both players believe that the other party wants to hurt them.

This model can provide with a sound account of some of the evidence presented above; in particular, it is obviously well-suited to explain the behaviour based on reciprocity that have been reviewed in 2.2.4, and it can also accommodate the attitude to punishment or revenge that is so frequently observed. According to Fehr and Schmidt (2001), this theory would be at odds with the presence of seemingly altruistic behaviour and concern for fairness reported in 2.2.2 and 2.2.3, as individuals can be co-operative even in the absence of a precedent kind behaviour by other agents. However, it has to be said that even this piece of evidence is not a direct contradiction of the theory. In fact, Fehr and Schmidt only consider a particular version this theory in which the *initial* expectations on others' behaviour do not count in terms of their choice. However, by supposing that people enter the experiment with a belief of a nice attitude of the other party to themselves, then even altruistic could be accounted for within a reciprocity model. Moreover, the decreasing degree of co-operation would prove that individuals adjust their beliefs as they observe less co-operation than initially expected or required, and this too would be consistent with the reciprocity hypothesis. Introducing *initial beliefs* as explanatory factors, which conforms to what is called 'norm of co-operation' hypothesis (see 2.2.4), could of course be seen as adding some arbitrariness in the theory. However, in principle it does not seem prohibitive to take explicitly account of such element and tested in experiments. I believe instead that the main shortcoming of the theory lies in the observation that people execute costly action even to 'protect' the interests of third parties. This shows that reciprocity can be referred to not only one's own payoffs, but also to that of others, which – as argued in 2.2.5 - seems to suggest that individuals take on an impersonal view when

assessing the attitude of other agents. It will be on this supposition that the model in the next section will be built.

2.4.2 Social Preferences

An alternative hypothesis on how other-regarding preferences enter individual motivations focuses on the final outcomes of the interactions rather than on the intentions of the players involved. In particular, individuals have 'social preferences' when their utility function depends, in some way, on the payoffs attributed to the other players as well as on their own. Such a definition is obviously general enough to encompass many –often-conflicting - specifications.

First, various forms of *altruism* fall into this category. These range from the more extreme version of *pure* altruism, in which an agent's utility *only* depends on the utility of some other agents, to the more general version in which the partial derivatives of one's utility with respect to the payoffs of *any* agent –including the subject herself- are strictly positive. Charness and Rabin (2000) considers a more elaborate version of altruism, which they call *quasi-maximin preferences*, where the other-regarding function is a convex combination of the Rawlsian maxinim criterion and of a utilitarian social welfare function.

That some altruistic traits plays a relevant part in at least some situations illustrated earlier has been repeatedly stressed (see section 2.2.2). However, it is equally clear that altruism cannot be considered a comprehensive account, as in some other situations intentions and reciprocity do matter, in particular when players punish others' behaviours perceived as 'unfair', despite the positive costs for themselves, as in the Ultimatum Game. This is the reason why Charness and Rabin extend their model to embed an element of reciprocity (see next section).

A different hypothesis on the way individuals look at payoffs allocation goes back at least to Veblen (1922), and argues that individuals, in addition to their own payoff, also care about their relative standing in the overall distribution. For instance, in the two-agent case, the utility function would have the following specification:

$$V_i(x_i, x_j) = h(x_i, x_i / x_j) \qquad (2.11)$$

where x_i and x_j are the payoffs obtained by the i-player and the j-player respectively, and the partial derivatives are positive with respect to both arguments. Notice that this specification implies the *opposite* of altruism, as utility diminishes when the other player's payoff increases. It is thus at odds with the other-oriented behaviour frequently observed in experiments.

Another account trying to combine the possibility of altruistic and spiteful behaviours by the same agent is that put forward by Levine (1998). In his model, a pair of parameters determines whether an agent is inclined to altruism or spitefulness, and the extent to which she wants to 'reciprocate' others' dispositions. More formally:

$$V_i = x_i + \sum_{j \neq i} x_j \left(a_i + \lambda a_j\right) / (1+\lambda) \qquad (2.12)$$

A positive (negative) sign of a_i implies that the agent has a disposition to be altruistic (spiteful); λ determines how the agent evaluates others' dispositions. Thus, spiteful agents will score less well into the subject's individual function, thus prompting a less benevolent attitude towards them. If $\left(a_i + \lambda a_j\right) \leq 0$, it could even be the case that an agent whose attitude is altruistic behaves spitefully against a spiteful individual. To be sure, an agent often cannot knows others' disposition. This makes the situation of repeated interaction analogous to a sequential game in which any agent tries to acquire information on the others' *types* on the basis of the actions executed in the previous rounds of the game. This account has the merit of conditioning the willingness to perform altruistic behaviour to the 'type' perceived in the other party, although how this happens is not formally specified. However, a serious limitation lies in that the disposition to being altruistic or spiteful is fixed and cannot change as an effect of the interaction and/or the types of players the subject is interacting with.

A more elaborate approach to social preferences comes from the so-called *inequity aversion* hypothesis. For instance, Fehr and Schmidt (1999) assume that *"a player is altruistic towards other players if their material payoffs are below an equitable benchmark, but she feels envy when the material payoff of the other players exceed this level"*. It is notable how this theory explicitly introduces a notion of an equitable distribution of payoffs as a benchmark to which agents attach intrinsic value. Assuming, as is likely to be the case

in many experimental contexts, that the egalitarian distribution is the one that individuals consider equitable, they put forward the following specification:

$$V_i(x_1,....,x_N) = x_i - \left(\frac{\alpha_i}{N-1}\right)\max\sum_{j\neq i}\{x_j - x_i, 0\} - \left(\frac{\beta_i}{N-1}\right)\max\sum_{j\neq i}\{x_i - x_j, 0\} \quad (2.13)$$

The additional hypothesis that $\alpha_i \leq \beta_i$ ensures that an agent is more concerned with her own relative position in the standing than with that of other agents. Moreover, the agent is altruistic only towards agents who are worse off then her in the payoffs standing. As a result, both positive *and* negative actions of an individual towards other players can be accommodated within this setting. This aspect certainly represent an important improvement with respect to the previous specifications based on social preferences. However, in section 2.2.4 it has already been argued that limiting the attention to the concern for fairness makes it impossible to account for the facts that are clearly related with the human attitude to reciprocate and base their actions on the perceived intentions of others. Moreover, as discussed earlier, it seems the case that the notion of fairness individuals draw on has trait of impersonality rather than being based on the standpoint of the self. This calls for an approach that somehow links the aspect of reciprocity with that of fairness concern.

2.4.3 Merging Intentions and Social Preferences

Some attempts in the direction just suggested already exist in the literature. For instance, in Charness and Rabin (1999) the 'weight' that each individual attaches to each other individual in her own social preference on the surplus distribution depends on the disesteem with which the agent herself thinks of the others, which is appraised in terms of the 'distance' of others' behaviour from a purely disinterested one. Likewise, in Falk and Fischbacher (1999) each agent computes a 'benevolence term' for any other agent, which depends on the degree to which any other agent's action has increased or diminished the inequality in the overall distribution. This term is then multiplied by a 'reciprocity term' that is positive or negative in relation to the other agent's action being perceived as kind or hostile. Finally, another parameter measures the relative weight attached to material utility with respect to that of reciprocity on the social distribution.

I believe that these models go in the right direction of coupling a concern for fairness with a reciprocity element. Nevertheless, I still believe that their account of fairness or social preference is still too restrictive and does not really capture what is in my view a key element in this component, i.e. the *impersonality* of the perspective that is adopted. In fact, the use of an inequity aversion and of a mixture of a Rawlsian and a utilitarian welfare function is not clearly motivated, and fails to satisfy this condition.

2.4.4 Normative Expectations

Another account, which has not received as much attention from the experimental point of view as the previous ones, is the theory of normative expectations. This neglect is somehow unjustified, though, as on the contrary the theory addresses some relevant questions still left unanswered in experimental economics. In particular, as I have tried to argue in the review of the evidence carried out in section 2.2 and in the assessment of the reciprocity hypothesis in section 2.4.1, the effect of initial beliefs on others' attitude towards the self may play a relevant part in the explanation of many experimental facts. As the theory of normative expectations can be seen as an attempt to address this aspect from a formal point of view, in that it provides a particular way to answer the question of how such initial beliefs are formed, it plays in fact a relevant part in the explanation of experimental evidence. Not only is this theory important for this reason, but also it plays a key role in the issue of 'stability' of social norms (see section 3.5 and Chapter 4). In what follows I shall review its underpinnings (sec. 2.4.4.A) and some related applications that have been proposed (sec. 2.4.4.B).

2.4.4.A The Resentment Hypothesis

In his *Theory of Moral Sentiments* (1759) Adam Smith emphasised the significant role of the expectations nurtured by members of a community in orientating one's behaviour:

"What reward is most proper for promoting the practise of truth, justice and humanity? The confidence, esteem and love of those we live with. Humanity does not desire to be great, but to be beloved." (Smith, 1759/1982, p. 166). *"We are pleased to think that we have rendered ourselves the natural objects of approbation, though no approbation should ever actually be bestowed upon us:*

and we are mortified to reflect that we have justly merited the blame of those we live with, though that sentiment should never actually be exerted against us"(Smith, 1759/1982, p. 116).

Sugden refers to this simple but fundamental assumption about human psychology the *resentment hypothesis* (Sugden (1998a: 16)). As I shall illustrate in the next section, an other-regarding motivation can be grounded on this condition and matched with a self-interested motive to shape an overall system of individual preferences. Expectations come to be considered 'normative' since they provide us with a strong sense of commitment to the pursuing of the rule generally followed in the community, conferring the character of obligation typical of norms. When a rule is established, each agent understands that its breaching would trigger a sense of resentment in other members of the community and as a consequence a sense of guilt in ourselves, thus urging us to refrain from flouting it (at least until the possibly contrasting self-interest becomes too strong). Expectations take on a cognitive aspect too. Indeed, agents perceive the norm by means of the set of expectations that all of the other agents of the community -the direct opponents of the interaction and the bystanders not directly involved in it- address to her. This introduces a further reason for which agents may want to abide by the rules of a community. For the norm may indicate what the public interest of a community is, thus inducing an individual to follow it because of her internal commitment to act in accordance with the general public interest, rather than to the resentment hypothesis (Pettit (1990: 731)). This element may be relevant in that sometimes the public interest is in contrast with the reigning norms, in those situations in which a norm is patently 'wrong' or inefficient. This argument brings on the question of the stability of a norm, and of its change – or revision – over time (Ulmann-Margalit (1990)), which will be analysed in depth in the next Chapters.

In addition, many scholars doubt that the account of cooperation in a Prisoner's Dilemma-like situation grounded on the idea of the adoption of a tit-for-tat strategy is actually successful. It has been pointed out that, especially in the many-players PD, the interaction may be such that the defection of one single agent may cause so negligible costs that the punishment from the rest of the agents may be economically inefficient. In these situations the tit-for-tat may not be a viable self-enforcing strategy to bring about an outcome of reciprocal cooperation (Pettit (1989a: 341-344); see also Brennan

and Pettit (2000)). Therefore, so the argument goes, an explanation based on normative expectations gains credibility in these contexts, since the costs implied by the mere observation of others' behaviour and the consequent sentiment of commendation or disapproval are virtually nil[10].

Pettit (1990: 742-745) elaborates on this argument to show how it is possible to account for the establishment of norms by means of an argument drawing on normative expectations – which he calls an *attitude-based derivation* – instead of the usual line of reasoning grounded on individual interests and reciprocal expectations, like for instance Lewis's – a *behaviour-based derivation in his words*[11]. He singles out five conditions for the proof to work:

(1) *Interaction assumption*, referring to the collectively beneficial character of the norm (see the previous section);

(2) *Publicity assumption*, stressing the necessity that the behaviour of the agents involved be observable by the others involved;

(3) *Perception assumption*, referring to the possibility for the agents to clearly make out whether everyone's action was for or against the benefit of the community;

(4) *Sanction assumption*, that underlines how the enforcement of a norm can be brought about by the attitude to encourage the obedience of the norm and discouraging its transgression embedded in agents' dispositions;

(5) *Motivation assumption*, equivalent with the resentment hypothesis.

These conditions shape what is the "natural" environment for a norm to come forward as a regularity of behaviour, and would have an explanatory function even when an account grounded on the behaviour-based strategy is not viable.

---

[10] However there exist other accounts of how a tit-for-tat story may be used in order to sustain a cooperative equilibrium: one of particular interest is put forward by Hardin (1988, p. 105) when he argues that this emerges from the adoption of tit-for-tatting in two-party PD, which then spreads to the whole population because of the tendency of individuals to employ the same norm in situations similar, though different, with respect to those previously met. For a similar account of how norms previously formed in small groups can then transmit to larger group in a sort of "contagion" see Bicchieri (1990: 855-861) and Anderlini and Ianni (1996). On how tit-for-tatting can be rational even in PD of finite length see Pettit and Sugden (1989b). For other accounts of the formation of cooperative behaviour though not based on tit-for-tatting, see Taylor (1987).

[11] One of the opponents of such an attitude-based derivation is Buchanan (1975: 132-33). He argues that such a strategy overlooks that both the component of discovering violators and punishing them

2.4.4.B Models of Normative Expectations and Social Status

Sugden's specification conforms to the idea of multiple reasons to action in individual preferences, where an other-regarding motivation adds to self-interest and the former builds on the resentment hypothesis and the idea of normative expectations. In one of his works (Sugden (1998a)), the formal device he constructs in order to give substantive content to the concern for normative expectations is what he calls an *impact function*, which depicts the effect of one's action on others' welfare given others' expectations. In other words, the main idea of Sugden is to associate the community expectations with the expected payoff of an agent.

In terms of the general version of the comprehensive utility function expressed in equation 2.5, we now have that the normative principle $T(\sigma)$ on which the other-regarding motivation is based, merely consists of the individual expected payoff by a player's opponent, who is labelled as *j*: $T(\sigma)=U_j(\sigma)$.

Some further caveats apply. First, I here deal with the simple case of a two-person stage game, where the two players are drawn at random from two different populations - say, *i*-players and *j*-players. We focus on the choice of a generic *i*-player, who then has to take into account the expectations on her behaviour formed by a *j*-player. Therefore, the generic profile of mixed strategies is substituted by the *percentage* of players playing each strategy in the population, which I label with p_i and p_j.
We now need to assess how *i*'s actual choice compares with what was expected from her. This is provided by the *impact function*, which is the difference from *j*'s *actual* payoff after *i*'s particular choice, and *j*'s expected *a-priori* payoff:

$$m(\sigma_i; p_i, p_j)=U_j(\sigma_i; p_j)-U_j(p_i; p_j) \qquad (2.14)$$

In this formula, σ_i represents the strategy *actually* played by *i*, whereas p_i stands for the average behaviour within the *i*-player population. Therefore, when $m(\sigma_i; p_i, p_j)<0$ the *i*-player is failing to conform to the norm as agent *j* experiences a payoff lower than expected. Conversely, if $m(\sigma_i; p_i, p_j)>0$ agent *i* is performing an action that

---

involve positive costs. But, as already stressed in the exposition, the enforcement would actually be costless if the resentment hypothesis and the concept of normative expectation held.

awards agent *j* with an extra-payoff than expected; in Pettit's words, *i* is performing a *super-erogatory* action. However, only the former of these two aspects is relevant for the resentment hypothesis. In fact, according to Sugden's formulation, an agent only suffers resentment when failing to live up to others' expectations, but she does not (necessarily) experience the contrasting sentiment of satisfaction when conforming to the convention to a degree even 'greater' than expected (Sugden (1998a: 27)). Consequently, the final form of the other-regarding source of utility is as follows:

$$g(T(\sigma);b)=\begin{cases}0 & \text{if } m(\sigma_i;p_i,p_j)\geq 0\\ m(\sigma_i;p_i,p_j) & \text{if } m(\sigma_i;p_i,p_j)<0\end{cases} \qquad (2.15)$$

Notice that the dependence on the difference between actual and expected payoff has here been assumed linear for simplicity, despite Sugden only constrains overall utility to be monotonically decreasing in *m* when this is negative. Sugden also provides a different model of normative expectations, which takes on a dynamical approach and is based on the analytical tool of *potential games*, where two agents are randomly matched to play a stage game from a population having continuously different types (Sugden, 1998b).

A different, but related, thread of literature is that in which agents have preferences for 'social status'. In Bernheim (1994), for instance, this is modelled as an additional component to self-interested utility, thus conforming to the general model represented in equation (2.5). The difference with Sugden (1998a) is in that disapproval is not attributed to possible losses in the payoffs with respect to what expected caused by the individual's action, but disapproval is directly assigned to the *action* of the individual itself. More precisely, supposing that different actions are somehow measurable on some scale, the higher the 'distance' of individual action from the average, the higher the loss in social status the agent will experience, the lower her comprehensive utility. In this framework, a static analysis of equilibrium brings about 'endogenous' formation of social norms, which change as the pattern of individual preferences changes. A norm can be said to emerge when individual behaviour by all agents cluster around a single action. However, equilibria emerge with 'multiple' norms, i.e. individual actions cluster around more than a single action. This happens when individual preferences are relatively 'dispersed' over the interval of possible actions, so

that what can be interpreted as anti-conformist behaviour becomes possible. A similar result of 'segregation' in different classes of behaviour, which also correspond with classes of income, is provided in Cole *et al.* (1998). Lindbeck *et al.* (1999) apply this framework to the study of the welfare system: it is shown how different patterns of social norms can exist, which makes more or less costly the decision to work as opposed to live off public transfers. Even in this case, social norms are determined endogenously, and multiple equilibria arise in response to changes in preferences and also in the size of the public transfer.

The next two Chapters will be devoted to an extensive analysis of the normative expectations idea in the account of social norms. This will be criticised on the grounds that normative expectations can only have a role in underpinning the *ex-post* stability of a norm, but cannot have heuristic value in an *ex-ante* perspective, i.e. in the explanation of their emergence. This further suggests that the theory should be further developed in order to have a significant role in the accounting of experimental evidence, as nearly every outcome could seemingly be sustained upon the choice of a suitable vector of initial beliefs.

2.4.5 Team reasoning

Another interesting theoretical approach to account for co-operation in strategic interactions is the so-called *team thinking* line of argument. Such an approach differs from those outlined above in that its aim is not so much the specification of a utility function describing the other-regarding motivations of individuals, but rather the attempt to offer a theoretical model to the idea that individuals come to share common objectives and act like 'teams', thus solving co-ordination problems, in many situations of strategic interaction[12]. The analysis of this theory is particularly interesting in the light of the analysis carries out in section 2.2, where the presence of forms of group-oriented behaviour has been repeatedly endorsed as a likely explanation of the facts under investigation. However, in the present section it will be argued that the

---

[12] The first models developed in this area were the so-called theory of teams (Marschak and Radner (1972)) and of multi-agent systems (Fagin *et al.* (1995)). It is notable that the general aim was to find the optimal action for the organizer of a team in a situation of uncertainty over the 'degree of

main theoretical tool on which this approach is based, i.e., that of team reasoning, does not seem to be consistent with the idea that individuals take account and trade-off between multiple reasons to action in their motivational system, which seems to emerge as a stylised fact from experimental evidence (see section 2.2.5). In other words, according to this view each agent is only a player of *her own* team, though her jersey may be a patchwork made up of the flags of many other teams. That is, an individual may follow different objectives, but this very fact makes her a player of one single team.

The main underpinnings of this approach are works carried out in the field of experimental and social psychology – e.g. Brewer and Kramer (1986); Brewer and Gardener (1996) – and by theorists of collective agency – e.g. Gilbert (1989); Tuomela (1995). The former emphasise the key importance of the collective dimension in individual agency. In the words of Brewer and Gardener (1996), a person has three radical ways of thinking of *herself*: as a "personal self", a "relational self" and a "collective self". This is what makes it possible group-identification, i.e. the perception by an individual of a tie, a sense of kinship or affinity, with a group of people, which often leads to the perception of a communality in their objectives. Such a sense of group-identity is typified by the use of expressions carrying the first person *plural*, rather than the *singular*, in many choices actually carried out by individuals.

The further question that is posed is how a group comes into being, and how its boundaries are set, even in the instances in which a group is not 'institutionalised' or coerced from outside, such as in the case of family relationships or hierarchical organisations. The answer is that, although in many cases it is possible to recognise an 'objective' concurrence of interests between some individuals, e.g. in co-ordination problems, that naturally helps to buttress the sense of belonging to a 'group', in other instances a team could all the same arise on the grounds of nothing more than the presence of some 'focal points' in the setting of some boundaries between some 'us' and some 'others'. This feature is all too well known to experimental economists, who repeatedly find out that even an accidental fact such as being drawn under the same

---

commitment' to the team of its various members, or their 'operative state' in case the members of the team are not human beings but machines subject to failure.

labelling in an experimental setting, triggers the development of team-oriented behaviour amongst those subjects.

The latter line of thought recognises the relevance of the 'we-thinking' at the psychological individual level, and tries to spell out its implications for the account of co-operation in collective actions problems. If this is inevitably made easier to account for, the problem then becomes to explain how the we-thinking line of reasoning is compatible with the *reductionist* strategy common to methodological individualism, i.e. the idea that *"every proposition about team agency or team preference is reducible to some definite propositions about individuals"* (Sugden (2000: 176)). After examining how the account offered by other theories is yet unsatisfactory, Sugden (2000) seems to offer the most convincing answer: 'team' preferences can be treated on a par with standard 'individual' preferences, in that the problem of the content of the former preferences is a matter of empirical investigation rather than rationality. In other words, one can take as granted the existence of a structure of 'revealed' team preferences in the same way as standard economic theory assumes that individual preferences are revealed in an ideal situation of choice. Once the existence of team preferences is accepted, the basic tenets of the reductionist approach can be said to be fulfilled. For single agents can engage in 'team-directed reasoning', i.e. they can carry out inferences about the group from the standpoint of individuals. In fact, it is only when team-directed reasoning is coupled with full *confidence* about other members' team-oriented reasoning and willingness to conform, that a notion of *team agency* can be said to arise.

The notion of team-directed reasoning is then crucial to a sound epistemological foundation of team-thinking. Bacharach (1999) offers the most general formal account of such a notion, which is interesting to analyse in order to shed more light on this theory. In fact, acting as a member of a team entails both the sharing of the same objective by all members, and team-reasoning. Bacharach allows for the possibility of every individual belonging to different teams, where the assignment to either of them is uncertain but regulated by an exogenous probability function that is common knowledge. Some key assumptions are that an individual can act for only one team, and that a team could be formed by a single individual on her own. Hence, once an individual knows to which team she belongs to, and has a probability function on

other agents' possible assignments, she computes the *vector* of actions maximising her own team objective function, taking into account the possibility of players lapsing for other teams and thus determining the (expected) optimal vector of strategies by other teams. After having determined such a vector of optimal strategies, the agent will perform the individual action prescribed by that vector. Bacharach then proceeds to find out the formal properties of the equilibria under team-reasoning; in particular he highlights the difference with respect to a seemingly related behaviour, that of so-called benefactors. Such would be the behaviour of agents whose individual objective function *coincides* with that of the team, but who do not team reason, that is, they do not solve the aggregate problem for the team to then derive their optimal individual strategy. The result is that team-oriented and benefactor behaviour do indeed differ, the equilibria associated with the former being a subset of those engendered by the latter. The intuition for this result is that when individuals team reason they take into account the behaviour of other members of their team, thus eliciting a somehow 'better', i.e. more advantageous in terms of their team obejectives, co-ordination amongst them.

Such a formal account of team-directed reasoning enables us to highlight some of its shortcomings, which justify the somehow different direction in modeling other-regarding motivations that will be taken in the next section. The major difference lies in that team reasoning does not conform to the account underlying equation (2.5). The reason is that, whereas Bacharach assumes that each individual is only assigned to *one* team at a time, in our model, apart from the case in which $\lambda_i=0$, the individual *simultaneously* belongs to *two* teams, i.e. that formed by all relevant players (second term of (2.5)) *and* the individual herself. Hence, the kind of 'trading off' between self-regarding and other-regarding motivations that is implicit in the multiple motivations approach to modeling preferences is no longer possible. In terms of Bacharach's terminology, the agent represented in (2.5) is a *partial* benefactor, in that she takes into account both the team objective function and her own personal preferences. The result is that such an actor *may well end up performing an action that does not maximizes neither objective of her 'teams'*, but realizes the *best compromise* between them. In other words, since (2.5) implies the possibility of a continuous trade off between self-regarding and other-

regarding utility, the weight λ_i being the rate of substitution between such sources of utilities, the action optimizing the comprehensive utility function could well be one that is not the best for either self-regarding or other-regarding objectives.

What is more relevant is that such an agent does *not* team reasons. For even when she considers her other-regarding preferences, she cannot take for granted a full compliance with the team objectives by other agents, as they too are partial benefactors. In terms of Sugden's account (2000: 194), an agent cannot have *full team confidence* about other potential members' compliance, because they too will perform an action that is not necessarily the best in terms of the team objective function. In fact, from the formal point of view of the game, being a partial benefactor is tantamount to belonging to a *single* team, i.e. that formed by the agent as the only member, although her objectives are shaped in accordance with the team objectives. But this implies that the setting is the standard one in game theory, i.e. agents perform the best action for their individual (comprehensive) utility function, given expectations of others' actions.

One can notice that the formal apparatus developed by Bacharach could easily accommodate for the case of multiple belonging to teams, where a probability could be assigned to any other agent being a partial benefactor as above and a full team member. One could substitute a *multi-valued* function of assignments of agent to team for the single-valued one that Bacharach adopts in his account, so as to make the possibility of multiple belonging formally treatable. When an individual were drawn in the 'multiple' team, then, he would be a partial benefactor for any team he belonged o, in the same fashion as in (2.5). In other words, there would be a vector of weights stating the degree of importance that the agent assigned to a particular team, relative to the others. Any team member could then work out what the optimal action for a partial benefactor is, and, on the basis of that, compute the best action for the team. But the main point is that such an account would be meaningful only under the assumption that, in the case that an agent is drawn to be a team member, there is full team confidence about her commitment to the team. As already pointed out, when any agent is a partial benefactor, such an assumption cannot hold.

However, even leaving aside the much greater analytical complexity that such an extension would require, I believe that this would constitute a significant departure

from the core of Bacharach's proposal. In fact, team reasoning calls for an individual solving the maximization problem for her team given the information on the team assignment probability distribution. This is made possible by the assumption that the action performed by a lapsing agent is known, so that the team objective function can be optimized in expected value. However, in the case that the lapsing agent belongs to more than one team, it may not be so clear-cut to anticipate what such an action could be, because

In conclusion, the idea of multiple belonging to teams seems to water down the notion of team reasoning. In fact, even admitting that a generic member of a team knew the preferences of all of the other players in terms of self-regarding and other-regarding preferences, any attempt to replicate their choices would be done considering them as *individuals*, rather than as member of teams. Although the idea that an agent complied with one team at a time can perhaps picture particular situations of interactions, in my view it still is a somehow implausible assumption for a general account of human behaviour. It is equivalent to saying that an individual commits to the prescriptions of a reason to action *at a time*, rather than considering *simultaneously*, weighing them up and then making a decision *all considered*. But this requires an agent to be a partial benefactor, not a team-player. It will then be along these lines that I will develop my analysis in the next section.

2.5 Other- Regarding Motivations as Conditional Compliance with a Shared Normative Principle

In this section I develop a model of individual choice that is consistent with the general setting illustrated in section 2.3, and that in my view offers a more comprehensive – and thus better - account of the empirical evidence illustrated in 2.2 than the contributions reviewed in section 2.3. Admittedly, its main limitation lies in that no specific normative criterion will be specified in this version. However, this is done on purpose in order to stress the generality of the model, and to emphasise the impersonal character of the normative criterion, in accordance with the evidence analysed earlier. A particular specification will be provided in the final section of the

Chapter. The focus is here on how agents balance self-interested and other-regarding motivations in a way that couples reciprocity and concern for morality. The resulting other-regarding motivation may be regarded as consisting of a *conditional* compliance with a general moral principle, where an agent construes the expected conformity of other agents as an incentive for her own conformity. In what follows, then, it will be assumed that agents share a moral principle that creates an ordering of the states of affairs in terms of their degree of fulfilment of the principle, on which they base their evaluation about one another's conformity. The issue of the convergence of individuals on a common moral principle relies on the inter-subjective character of moral judgments, which has been illustrated in section 1.4.

2.5.1 Compliance with Morality Based on Reciprocity

The notion of reciprocity that we shall develop builds on Rabin's idea (see section 2.4.1), with the basic difference that mutuality is not directed towards the agents' own material payoffs, but is appraised with respect to the compliance with the set of normative principles that are considered by the individuals. Therefore, Rabin's pair of kindness functions (2.7) and (2.8) is substituted by functions of compliance with the normative criterion, in such a way that each agent's incentive to perform an action satisfying the moral criterion, and possibly contrasting the self-interested reason to action, is positively linked with the extent to which the opponent is performing an action consistent with the same normative criterion[13].

The model that I wish to develop emphasises the aspects of reciprocity in acting in accordance to the shared ideology, as represented by the normative function. The idea I want to pursue is one really widespread in the literature on morality, namely, that agents are available to sustain an action beneficial to some 'objective' idea of public interest, though possibly detrimental in terms of self-interest, only if they expect the same behaviour from other agents of their group.

We now restrict our attention to a two-person game, even though a generalisation to the case of n players would be straightforward. In analogy with Rabin's pair of

---

[13] An interpretation of this other-regarding motivation in terms of *deontological* reasons to action (see section 1.2.2) can be found in Sacconi (2002) and Sacconi and Grimalda (2002).

kindness functions, measuring the mutual impact of one's actions on the other's individual utility, we can now introduce functions computing the degree of compliance to the normative principle. We first define *i*'s compliance to this in the following way:

$$f_i\left(\sigma_i, b_i^1\right) = \frac{T\left(\sigma_i, b_i^1\right) - T^{MAX}\left(b_i^1\right)}{T^{MAX}\left(b_i^1\right) - T^{MIN}\left(b_i^1\right)} \tag{2.16}$$

where $T^{MAX}\left(b_i^1\right) = \arg\max_{\Sigma_i} T\left(\sigma_i, b_i^1\right)$ and $T^{MIN}\left(b_i^1\right) = \arg\min_{\Sigma_i} T\left(\sigma_i, b_i^1\right)$. In other words, $T^{MAX}\left(b_i^1\right)$ and $T^{MIN}\left(b_i^1\right)$ represents respectively the maximum and minimum value that the social function can assume, depending on *i*'s action, given *i*'s first order belief b_i^1 over the action that *j* is going to perform[14]. Therefore, if $T^{MAX}\left(b_i^1\right)$ $\left(T^{MIN}\left(b_i^1\right)\right)$ is obtained, then agent *i* is maximising (minimising) the normative function given his first order belief. $T\left(\sigma_i, b_i^1\right)$ is instead the value of the normative function corresponding to *i*'s actual choice σ_i.

Hence, $f_i\left(\sigma_i, b_i^1\right)$ is an index varying between -1 and 0 expressing the extent to which *i*'s action satisfies the normative criterion associated with the function *T*. When $f_i\left(\sigma_i, b_i^1\right)$ is equal to 0 (-1) it means that *i* is exactly performing the strategy maximising (minimising) the normative function, given *i*'s first order belief, and this testifies that his action is consistent with the normative prescriptions at the maximum (minimum) degree. For instance, if the normative principles coincide with the moral principles, the compliance with morality is measured by the extent to which one's action increases the social function *T*.

To model the concept of reciprocity in the individual motivational system, we need to introduce a function symmetric to that set out above. This is the esteem that player *i* forms about *j*'s compliance with the normative principle:

$$\tilde{f}_j\left(b_i^1, b_i^2\right) = \frac{T\left(b_i^1, b_i^2\right) - T^{MAX}\left(b_i^2\right)}{T^{MAX}\left(b_i^2\right) - T^{MIN}\left(b_i^2\right)} \tag{2.17}$$

---

[14] Notice the dependence of $T^{MAX}\left(b_i^1\right)$ and $T^{MIN}\left(b_i^1\right)$ on the belief b_i^1. Indeed the belief is necessary in order to determine the probabilities for the expected value of the welfare function, which is:

where $T^{MAX}\left(b_i^2\right)=\arg\max_{\Sigma_j} T\left(b_i^2,\sigma_j\right)$ and $T^{MIN}\left(b_i^2\right)=\arg\min_{\Sigma_j} T\left(b_i^2,\sigma_j\right)$. Therefore, $T^{MAX}\left(b_i^2\right)$ and $T^{MIN}\left(b_i^2\right)$ represent the value that the social function takes when player *j* respectively maximises or minimises it, given the second order belief of player *i*. In other words, those functions indicate the maximum and minimum values that player *j* can attribute to the social function, given the belief he has about *i*'s action as perceived by *i* himself. In fact, recall that such a function measures the *esteem* of *j*'s compliance to the normative principle as measured from *i*'s standpoint. Thus, if player *i* has formed a belief b_i^2 about the player *j*'s belief over *i*'s action, he will judge *j*'s actions from this point of view. He will then consider the best and worst value that *j* can do with respect to the normative function, and then compare these values with $T\left(b_i^1,b_i^2\right)$, which is the actual value that *i* expects the social function to get according to his beliefs. Alike the twin function $f_i\left(\sigma_i,b_i^1\right)$, a value of $\tilde{f}_j\left(b_i^1,b_i^2\right)$ equal to 0 (-1) indicates the maximum (minimum) degree of compliance by player *j* to the norm as embodied in the social function *T*.

2.5.2 The Comprehensive Utility Function

We can now introduce the final version of the utility functions. Notice that, as in every psychological game, the utility of an agent depends on her beliefs over the different possible outcomes. We assume the following representation, which blends the two functions of compliance to the norm:

$$V_i\left(\sigma_i,b_i^1,b_i^2\right)=U_i\left(\sigma_i,b_i^1\right)+\lambda_i\left[1+\tilde{f}_j\left(b_i^2,b_i^1\right)\right]\left[1+f_i\left(\sigma_i,b_i^1\right)\right] \qquad (2.18)$$

The fact that b_i^1 now substitutes σ_j depends on the fact that only in equilibrium the two are assumed to coincide. The other regarding utility, again weighted by the coefficient λ_i, consists of the product of the two compliance functions augmented by 1.

$T\left(\sigma_i,b_i^1\right)=\sum_{S_i}\sum_{S_j}\overline{T}\left(s_i,s_j\right)P_{\sigma_i}\left(s_i\right)P_{b_i^1}\left(s_j\right)$, where the probability $P_{\sigma_i}\left(s_i\right)$ is what prescribed by the mixed strategy σ_i, and $P_{b_i}\left(s_j\right)$ is the probability computed in accordance to the formula of the previous note.

The idea we wish to capture through this specification is twofold. On the one hand, the agent's utility depends positively on the realisation of the 'best' social outcomes, in terms of the satisfaction of the normative criterion; indeed, the other regarding utility is increased when an agent performs an action increasing the value of T, whoever she is. The second aspect captured by this specification is that call of 'reciprocity', or of conditional conformity, in the compliance with the normative criterion: in fact, the (esteemed) compliance of the other player, as expressed by $\tilde{f}_j\left(b_i^1, b_i^2\right)$, may be seen as the 'marginal incentive' that the subject has in pursuing her other regarding motivations, as represented by $f_i\left(\sigma_i, b_i^1\right)$. Therefore, the other regarding utility grows as the counterpart's action is perceived as more consistent with the ideology, thus eliciting a similar behaviour in the agent too. In the extreme case in which $\tilde{f}_j\left(b_i^1, b_i^2\right)$ is equal to –1, which denotes the worst action that agent *j* can perform in terms of the normative criterion, the coefficient of the ideological motive gets equal to zero, thus leaving the self-interest as the only relevant motive to action[15]. Conversely, when $1+\tilde{f}_j\left(b_i^1, b_i^2\right)$ is positive and sufficiently "large", then agent *i* may accept to pursue an action that is contrary to her self-interest but conform to the normative criterion[16]. In general, the evaluation of the opponent's compliance with the

---

[15] To be sure, even when agent *i* performs her worst action in terms of self-interest, the fact that agent *j* acts contrarily or in favour of public interest does not affect *i*'s overall utility function. Therefore, we can interpret the situation where one or both the agent perform the action leading to the worst outcome in the welfare function as one in which the social contract between the agents breaks down.

[16] Of course this is only one of the possible models of the ideological motive to action. Another one, which, *mutatis mutandis,* coincides with Rabin's specification is the following:

$$V_i\left(\sigma_i, b_i^1, b_i^2\right) = U_i\left(\sigma_i, b_i^1\right) + \lambda_i\left[\frac{1}{2} + \tilde{f}_j\left(b_i^2, b_i^1\right)\right]\left[\frac{3}{2} + f_i\left(\sigma_i, b_i^1\right)\right]$$

An "equitable" payoff in the normative function is here identified with half of the difference between T^{MAX} and T^{MIN}. Hence, agent *i* will experience a positive incentive to perform an action increasing the social welfare only when the opponent performs an action above this level. However, if agent *j* executes an action below this equitable level, then agent *i* would be subjected to an incentive to act *contrarily* to the normative criterion. This specification seems to emphasise the aspect of reciprocity *per se* partly neglecting the other aspect of the will to contribute to the public interest. We have thought, though, that this emphasis was somehow inappropriate in the present context, thus opting for a specification in which the incentive provided by the opponent in acting for the public interest was always non-negative, and nil only in the extreme case of him inflicting the least value to the social welfare function. The first account captures the idea that agents are interested in the social outcome fulfilling the normative prescriptions without any concern for the other agents' commitment to the same principles. This specification can be taken as a useful reference point with respect to the more

normative criterion magnifies or shrinks the individual motivation to act in accordance with the normative principle as well.

2.6 An Application: The Constitution of the Non-profit Enterprise

It is the purpose of this section to provide a straightforward application of the model of choice set out in the previous section to the explanation of the non-profit organisation. The interest of this application lies in that a number of explanations (Rose-Ackerman (1996)) emphasise the presence of other-regarding motivations in the actors of the non-profit, which generally simply boil down to mere altruism. I believe that the model of individual behaviour developed so far, relying on a more sophisticated analysis of individual motivations, can offer a better account than those provided, also helping to shed some lights on some facts, like the frequent practise of self-imposing norms involving fiduciary duties and codes of conduct, which would be partly unaccountable for within the received literature[17]. First, I depict a situation of interaction in the production of a good, whose outcome is a variety of different organisational forms of a firm. This game is analysed in accordance with the two attitudes that make up the utility functions of the players (section 2.3). In section 2.6.2, the Nash social normative function is adopted as the normative criterion used by the agents, and we analyse the conditions under which a non-profit organisational form can be an equilibrium of the game. Finally, section 2.6.3 presents the equilibria for the game.

2.6.1 The Game of Production

I suppose that three players are involved in the game of production: a worker (W), an entrepreneur (E) and a consumer (C)[18]. The latter is actually a dummy player, her

---

elaborated version of the next section, other than an interesting account of the ideal motive to action *per se*.
A specification in which there was no concern for the agent with the action of the counterpart would be the following:

$$V_i(\sigma) = U_i(\sigma) + \lambda_i[T(\sigma)]$$

[17] For the complete illustration of the model, see Grimalda and Sacconi (2002).
[18] See note 1 in the first Chapter for the indication about the gender of the players.

actions not having any impact on the others' payoffs, though her payoff is affected by the others' actions. The worker and the entrepreneur work together in a firm, and are to decide the degree of their commitment to the company, which is supposed to be measurable along some scale. This in turn brings about different organisational forms for the firm. More specifically, each of the active agents has two available strategies, one prescribing the supply of an action that would be standard in a free-market, profit-oriented context. The other action permits the improvement of the quality of the supplied good, in exchange of an extra-cost with respect to the alternative strategy.

For instance, the entrepreneur may decide to adopt a productive practise, or a technology, which permits to increase the quality of the good, where, though, this technology is more costly with respect to that adopted in a purely competitive context. Analogously, the entrepreneur may renounce to a part –or all- of his profits in order to reinvest them in the productive process either by improving the quality or increasing the quantity of the good supplied at the same price. I shall indicate with h_E and l_E the adoption of the good's quality-improving action and that leaving the quality of the good unaltered with respect to market standards respectively, where the letters h and l refer to the high or low quality-enhancing purpose of the action, and the subscript E stands for the entrepreneur.

Analogously, the worker may decide to work for a wage inferior to that fixed in a free-market context, thus partially – or totally - supplying his labour contribution in a *voluntary* form. Similarly, he may increase his effort in the provision of the good at the same wage. In both cases, either the quality of the good is improved, or this is offered in a larger amount at the same price. I shall indicate this pair of action with h_W and l_W. The consumer does not have actions affecting the utility of the other two agents, but the surplus derived from the consumption of the good depends on its quality, thus on the level of effort put in by the producers.

Following the formalisation introduced in section 2.3, I distinguish between the set $I=\{W,E\}$ of the active players and the set $I^*=\{W,E,C\}$ that includes the dummy player C. A strategy set for the two agents can be easily introduced by considering that both have an action that improves the quality of the good and another that leaves it unaltered with respect to a competitive context. I indicate this with $S_i = \{hi, li\}$, $i \in I$.

Also recall that $S = \underset{i \in I}{\times} S_i$, where the generic element $s \in S$ indicates a vector of pure strategies for the two players, and that $\Sigma := \underset{i \in I}{\times} \Sigma_i$ is the set of mixed strategies profile, with generic element $\sigma \in \Sigma$.

The game representing the interaction depicted so far is then as follows:

	h_E	l_E
h_W	$\underline{w}, R - \underline{w} - c, s$	$\underline{w}, R - \underline{w}, \frac{s}{2}$
l_W	$\overline{w}, R - \overline{w} - c, \frac{s}{2}$	$\overline{w}, R - \overline{w}, 0$

Figure 2.1

The first, second and third terms in each box represent the material payoff for the worker, the entrepreneur and the consumer respectively. c stands for the extra cost that must be paid for by the entrepreneur if she wants to engage in the quality enhancing action of the good, namely h_E. R indicates the revenues of the selling of the good, which is assumed to be constant in all of the four possible outcomes, and w is the wage, which enters as a cost for the entrepreneur and as the only source of self-interested utility for the worker[19]. There are two possible levels of the wage: $\overline{w}$ is a comparatively high level that obtains when the worker supplies a level of labour in accordance with a market standard (strategy l_W), whereas $\underline{w}$ is a lower level that the worker is available to earn when engaged in the good's quality enhancing action (strategy h_W). Therefore, the difference between $\overline{w}$ and $\underline{w}$ is the cut in his real wage that the worker is available to accept in order to improve the quality of the good.

The consumer's utility is given by the surplus gained in the four possible outcomes. This depends on the effort put in by the other agents in improving the quality of the good. In particular, I normalise to 0 her level of surplus in the outcome where neither the worker nor the entrepreneur engage in the quality improving action, that is (l_W, l_E).

---

[19] For simplicity the self-interested utility of both worker and entrepreneur is assumed to be linear in the monetary revenue.

I then assume that when both agents agree to enhance the quality of the good, the surplus gained by the consumer is comparatively higher, equal with the level *s*, whereas when only one of the two agents contributing to production provides such an activity the surplus reaches an intermediate level, for simplicity equal with *s/2*.

I identify the outcome in which both agents perform the quality-improving actions as that leading to the constitution of a non-profit venture. The intuition is quite simple: provided that by construction the outcome (l_W, l_E) is associated with the level of effort supplied in a free market context, (h_W, h_E) takes on all the relevant characteristic of a non-profit-oriented firm, that is the entrepreneur gives up her profits to invest in a quality-enhancing technology, or simply to increase the quality or the quantity of the good, while the worker supplies a larger amount of effort or some voluntary work. The surplus of the consumer is then as high as possible. The other pair of outcomes represent different situations: (h_W, l_E) gives the best payoff for the entrepreneur as she can count on the worker giving the maximum of his effort while not performing any quality-increasing action; conversely (l_W, h_E) provides the worst payoff for the entrepreneur as the extra-costs that she sustains cannot be compensated by the provision of some extra-work by the worker.

If the game is played by the two active players with no concern for the dummy player and with regard only to their self-interested utility, then a unique Nash equilibrium in dominant strategies exists, in which both agents perform the low-quality action. However, the presence of some other-regarding motivations can overturn this result. This is the argument of the next section.

2.6.2 Contractarianism and the Ideology of the Non-Profit Sector

As already pointed out, the set of normative criteria moulding the conformist motive to action have not been attributed a specific shape yet. In fact, to the purpose of building up a model of choice, my main point was to emphasise the existence of a prompt to action differing from the self-interested one, emphasising the conditional willingness of the agents to abide by some general moral or ideological principle. In the present section, I shall assign the Nash social normative function the role of such a

general normative principle. Indeed, this function fulfils all the relevant properties of a contractarian view of morality, which, in my opinion, plays a fundamental role in the ideology thriving in the non-profit sector.

In fact, I believe that the philosophical goal of the non-profit venture may be thought of as that of *including* all the various categories of stakeholders affected by the production of a good as relevant parties in the definition of the choices and of the behaviour of the firm. In other words, the core of the "mission" of the non-profit company would be to take on the interests of those subjects who do not have a "voice" or a direct influence on the decisions of the firm, and act *as if* they indeed had a role in that. This is tantamount to saying that the non-profit firm would behave in accordance with a *hypothetical* contract regulating the terms of the productive activity and the distribution of the surplus[20].

2.6.3 The Selection of the Non-Profit Enterprise as an Equilibrium

Recall that the expression of the Nash normative function is as follows:

$$N(U_1,...U_N)=\prod_{i=1}^{N}(U_i-d_i) \qquad (2.19)$$

where d_i represents the reservation utility that agents can get when the process of bargaining breaks down, that is when they renounce to act in mutual cooperation. In the present context, I think appropriate to set all of these reservation utility to the level of zero. This is actually something more than an arbitrary choice, whose extensive justification would however lie beyond the scope of this section[21].

Applying this function to the present model, and expressing it with respect to the pair of the relevant agents' actions, I obtain the following values:

---

[20] On this view of the firm, see Sacconi (2000). For a review of the philosophical and cognitive properties of the Nash social welfare function, see Brock (1979) and Sacconi (2000).

[21] Many authors would argue that the proper choice for the "exit option" would be the Nash solution of the material game played non-cooperatively. However, this choice is not immune for criticism as a possible situation of prevarication of one party over the other in the *status quo* would carry over to the final "moral" solution. This is the reason why other authors have proposed the notion of a "moralised" status quo, in which some minimal form of reciprocal respect are already in place. Therefore, one may consider our choice equivalent with a, perhaps naive, notion of moralisation of the status quo from which the "bargaining" starts.

$$
\begin{aligned}
N_{hh} &\equiv N(h_W, h_E) = \underline{w}(R - \underline{w} - c)s \\
N_{hl} &\equiv N(h_W, l_E) = \underline{w}(R - \underline{w})\frac{s}{2} \\
N_{lh} &\equiv N(l_W, h_E) = \overline{w}(R - \overline{w} - c)\frac{s}{2} \\
N_{ll} &\equiv N(l_W, l_E) = 0
\end{aligned}
\tag{2.20}
$$

For a significant set of the parameters, I can assume that the Nash function is maximised in (h_W, h_E)[22]. Recalling what set out in the previous section, this would be the allocation obtained in the process of bargaining between the three agents.

It is now straightforward to show how this outcome can be viewed as optimal by the agents when the conformist utility is sufficiently high with respect to the self-interested. Specifically, I want to prove that (h_W, h_E) can be sustained as a Nash psychological equilibrium, as defined in section 2.2.3. Let us first consider the position of the worker and compute his level of utility associated with such an outcome. His self-interested utility is clearly the lower wage; what about his conformist utility? Recalling the expressions of the two functions measuring the compliance with the ideology, I can notice that, provided that N_{hh} is the maximum for the function, both compliance functions will be equal to zero, thus attributing the maximum value to the ideological source of utility: $V_W\left(h_W, b_W^1 = h_E, b_W^2 = h_W\right) = \underline{w} + \lambda$. Notice that in the computation of this value I have used the definition of the Nash psychological game equilibrium, which implies that the beliefs of the agents must be confirmed by the agents' actual choice. Accordingly, the beliefs assign probability one to the equilibrium strategies.

Let us now test whether the worker finds this allocation optimal or he has any incentive to deviate. In Psychological Games, a deviation from a certain allocation consists of a change in the agent's strategy, *given* the set of beliefs held in that allocation. In other words, when deviating, the agent must take into account what the

---

[22] In particular, $N_{hh} > N_{hl} \Leftrightarrow R - \underline{w} > 2c$ and $N_{hh} > N_{lh} \Leftrightarrow 2\frac{R - \underline{w} - c}{R - w - c} > \frac{\overline{w}}{\underline{w}}$. The first condition implies that the extra cost required for the quality improving technology is not too large in comparison with the profits of the firm when the worker accepts the lower wage. The second condition ensures that the increase in the consumer and entrepreneur's utility when the worker partly acts voluntarily compensates the loss in the earnings of the worker himself.

expectations of the other agents on his behaviour are, and then compute the possible change in his own comprehensive utility deriving from not conforming to such expectations. In our case, I shall generically indicate with $\sigma_W<1$ the probability with which the worker plays h_W in the mixed strategy adopted in the deviation. The estimation of the entrepreneur's compliance to the ideology is unaffected by this deviation, since by construction the worker knows that she still believes that he is going to perform h_W.

However, the worker's very compliance to the normative principle must change. Given that the entrepreneur is still going to perform with probability one h_E, the resulting value for the Nash function is: $N(\sigma_W, h_E) = \sigma_W N_{hh} + (1-\sigma_W) N_{lh}$. Given the worker's belief, his action that maximises (minimises) the Nash function is to play h_W (l_W). Formally: $N^{MAX}(b_W^1 = h_E) = N_{hh}$, and $N^{MIN}(b_W^1 = h_E) = N_{lh}$. Substituting these values into the function measuring the compliance of the worker with the normative principle, I obtain:

$$f_W(\sigma_W, b_W^1 = h_E) = \frac{(1-\sigma_W)(N_{lh} - N_{hh})}{N_{hh} - N_{lh}} = -(1-\sigma_W) \qquad (2.21)$$

Hence, the comprehensive utility of the deviation is:

$$V_W(\sigma_W, b_W^1 = h_E, b_W^2 = h_W) = \sigma_W \underline{w} + (1-\sigma_W)\overline{w} + \lambda\sigma_W \qquad (2.22)$$

The other regarding source of utility is now smaller: the worker is paying the fact that he is not reciprocating the action of the counterpart. Knowing that the entrepreneur is dong her best to act in accordance with the normative principle, the fact that he is partly failing in doing the same causes a lesser satisfaction deriving from the conformist motive. A different but related interpretation is that the worker feels guilty for not having conformed to the counterpart's expectations. On the other hand, the expected value from the self-interested utility is certainly higher. To ensure the optimality of the choice of the quality improving action for the worker, I therefore need a further condition:

$$V_W(h_W, b_W^1 = h_E, b_W^2 = h_W) > V_W(\sigma_W, b_W^1 = h_E, b_W^2 = h_W) \Leftrightarrow \lambda_W > \overline{w} - \underline{w} \qquad (2.23)$$

This condition states that the weight attributed to the ideological source of utility must be sufficiently large so to compensate the loss in self-interested utility caused by not performing the best action in terms of self-interest.

An analogous condition ensuring the pursuing of the quality improving action holds for the entrepreneur:

$$V_E\left(h_E, b_E^1 = h_W, b_E^2 = h_E\right) > V_E\left(\sigma_E, b_E^1 = h_W, b_E^2 = h_E\right) \Leftrightarrow \lambda_E > c \qquad (2.24)$$

I therefore have a simple intuition of how the presence of a conformist motivation in the individual system of preferences helps to support the selection of the equilibrium associated with the non-profit form of organisation. When the importance attributed to this is sufficiently high in comparison with the material gain that must be given up when acting in conformity with the normative principle, than the outcome in which both agents perform their best action in terms of the interests of the third party involved in the interaction, going against what the pursue of their mere self interest would prescribe. Hence, the presence of two agents motivated to act in accordance with the normative principles reigning in a society, which in turn shape a peculiar ideology in the non-profit firm, emerges as a necessary condition for a non-profit firm to be founded.

Up to now this result seems fairly natural: whenever two agents are sufficiently concerned with the compliance with the normative criterion, and when they have formed reciprocal expectations on that both will abide by such a criterion, then an conformist equilibrium emerges as a solution of the game. However, this is not the only equilibrium of the game: in fact, as frequently occurs in reciprocity-based models, the standard Nash equilibrium is also a Psychological Nash equilibrium. The intuition is indeed quite simple: if one agent's counterpart does not act in accordance with the ideological principle, then the agent herself has no incentive to do so; as a result, the only possible equilibrium is that where each agent cares about her self-interest, thus the Nash equilibrium, which is unique in the game is selected. In fact, the situation is similar to a coordination problem, as Figure 2.2, which depicts each agent's best reply function, shows.

σ_i, $i = W,E$ represent the probability with which h_i is played in the players' mixed strategies. It is worth noticing that there exists a threshold level in the best reply

functions such that each agent performs the 'good' action only if the action of the counterpart is sufficiently 'good' and *vice versa.* This gives rise to a third equilibrium, this time in mixed strategies, for the game.

Therefore, the presence of a significant attitude by the agents to perform the actions prescribed by the moral principle to a full extent is a necessary condition in order that the non-profit organisational form be derived as an equilibrium of the game. However, this condition is not sufficient: even when agents assign a large "weight" to their conformist motive to action, a failure in signalling their attitude to their counterpart may lead to the selection of the forprofit organisational form as the equilibrium. In Sacconi and Grimalda (2002) it is argued that codes of ethics can be used as cognitive devices to solve the possible co-ordination failure.

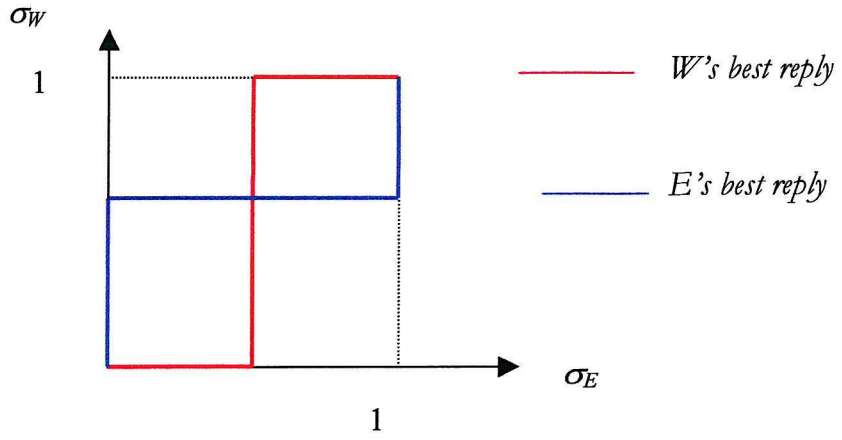

Figure 2.2

CHAPTER 3

INDIVIDUAL CHOICE WITHIN SOCIAL NORMS

INTRODUCTION

As I have outlined in the introduction, the relationship between norms and individual choice can be reversed with respect to the argument of Chapter 2. In the present Chapter, thereby, the focus will be on how social norms are determined through interactions amongst individuals, rather than on the aspect of their motivational spur on individuals.

Such a different perspective enables us to obtain a deeper understanding of the concept of social norm, since, as it will soon become clear, different notions of norms can be defined in accordance with the type of commitment required to the individual in complying with them. Hence, the type of motivations required to sustain a social norm will be used as the relevant criterion of classification. In particular, relying on the distinction between *self-regarding* and *other-regarding* reasons to action (see sections 1.2 and 2.1), three different notions of norms will be presented. Firstly, *mutually beneficial 'conventions'* are analysed (section 3.1), where not only does the action maximise one's self-interest but also that of the other individuals involved in the interaction. Secondly, *individually beneficial conventions* are characterised by the fact that the action of some individual is still the best in terms of her self-interest but it fails to optimise others' self-interest (section 3.2). The third notion is that of *other-regarding conventions*, where the action of some individual goes against her own self-interested behaviour as it is grounded on some types of other-regarding motivations (section 3.3).

In the classification that I suggest, each of these regularities of behaviour can be thought of as a *social norm*, as they all satisfy the requirements of the relevant

definitions put forward in the literature (see in particular Lewis (1969); Sugden (1986); Sugden (1998b)). Moreover, each notion can be matched with a different concept of 'solution' of the corresponding game: so, mutually beneficial Nash equilibrium are the relevant concept of solution for the first type of social norm, the standard Nash equilibrium is applied to the second, whereas the Psychological Nash equilibrium is associated with the latter.

In section 3.4, relying on Lewis's approach to defining conventions, the structure of actions and expectations underlying a norm is analysed; in particular, it is argued that the *cognitive* aspect relative to the structure of expectations on each other's behaviour and the *strategic* aspect relative to the choice of the 'optimal' action are interrelated. This analysis is used to develop an insight into the question of individual compliance with norms, and the question of their *moral* content is also investigated. Section 3.5 builds on this analysis in order to shed some light on the cognitive and strategic structure underlying the idea of normative expectations. I put forward a distinction between 'empirical' and 'causal' expectations, where the former rely on the force that past interactions have in shaping current expectations, whereas the latter require a justification of the action based on some independent, i.e. not determined by the relation itself, criterion of assessment. I argue that social norms based 'exclusively' on empirical expectations, which constitutes the typical example of normative expectations, are more 'fragile' than those grounded on causal expectations. For the cognitive and strategic structure underlying them is similar to that of conformative behaviour, thus they are likely to be 'unstable' with respect to even marginal shifts in aggregate behaviour.

Section 3.7 brings these considerations a step further by arguing for the necessity of a *dynamical* analysis of social norms. Hence, the Evolutionary Game Theory approach is presented as a response to this stance, and it is shown how this model leads to spontaneous emergence of rules of behaviour. It is on such instruments that the analysis that will be provided in Chapter 4 rests upon.

Before starting off with this analysis, though, let me be clear on some definitional aspects that may engender some confusion to the reader. Some authors seem to attach

a much sharper distinction between 'norms' and 'conventions' than what I do here[1]. The difference would lie in that, while in conventions the presence of more than one 'alternative' to the role of convention is needed, this is not necessarily the case for norms: for instance, mutual co-operation in a Prisoner's Dilemma stands out as a the only Pareto-dominant outcome, thus its emerging as a norm does not have a proper 'conventional' character in that there is no comparable alternative. Second, norms typically carry with them a sense of 'obligation' that, while representing their key aspect, is only an ancillary property for a convention. In fact, though conventions can be seen as a particular kind of norms, endowed with moral content, the sense of obligation that may assume is not necessary for their enforcement (see Lewis (1969)).

However, I will here follow the line of argument now common to some authors (especially Sugden (1998a); (2000a)), who think that the first distinction is after all not necessary, as, given the change in perspective provided by Psychological games, it is almost always possible to find multiple equilibria in situations that, were they analysed with standard game-theoretical tools, would only present a single equilibrium. Instead, the second aspect relative to the sense of obligation in furthering a norm will remain a key element in the distinction that will be proposed in this Chapter. Hence, norms will be here understood as a general category that encompasses the three different notions of conventions mentioned above.

---

[1] For instance Elster (1989) argues that norms should be considered as a different category than conventions, because of the different practical consequences that their breaching would bring about: material consequences would only be caused by infringing conventions, whereas breaking norms would only call for social 'disapproval'. Ignoring the question whether he is right in thinking in this way, the relevant aspect for our purpose is that he does not seem possible the overlap between norms and conventions, as well as between other patterns of behaviour. In particular, he draws a difference between *social* and *moral* norms, by defining the former as *"nonconsequentialist obligations and interdictions, from which permissions can be derived"*, whereas the latter varies according to the moral theory taken as a reference. In the case of utilitarianism, these could be thought of as *consequentialist obligations and interdictions* (Elster, 1990: 864). Moral norms would differ from *legal* norms too, as in the latter the self-interested motive of avoiding the punishment would be prevalent. In the approach I will follow, instead, social and moral norms are not conceptually separated, and they are seen as outcome-oriented. Moreover, the majority of scholars on the subject seem to view norms as peculiar institutions that reinforce or overlap with other patterns of behaviour, like customs, laws or social standards.

3.1 MUTUALLY BENEFICIAL CONVENTIONS

3.1.1 Co-ordination Problems

The first account of norm I will present can be traced back to the well-known Lewis's seminal contribution (1969). Although Lewis limits his enquiry to a fairly limited types of interactions, namely, so-called co-ordination problems, the insights he provides and the issues he tackles set the ground for the analysis that will be developed afterwards. Despite Lewis generically refers to conventions in his investigation, his notion corresponds with that of mutually beneficial conventions in the classification I propose.

A co-ordination problem is made up of two basic characteristics: first, agents have a *predominant*, if not perfect, coincidence of interests; second, at least two Nash equilibria exist in the game. This pair of features can be summed up in the requirement that two *mutually beneficial* Nash equilibria exist in the game; that is, the Nash equilibria are such that not only does the agent maximise her own payoff, but also she maximises any other player's payoff given the other players' strategies by performing the equilibrium action. In slightly more formal terms, a situation of predominant coincidence of interests occurs when, for every agent, given others' actions:

- anyone's strategy is a (strict) best reply to others' ones ((strict) *Nash* condition)[2];
- anyone's action maximizes anyone else's payoff (*mutual benefit* condition).

Such equilibrium is also referred to as a *(strict) co-ordination equilibrium.*

A situation of *pure* coincidence of interests obtains in symmetrical games, where agents' payoffs, possibly after suitable linear rescaling, are equal in every square; for instance, a *pure co-ordination game* is as follows:

	C1	C2
R1	1,1	0,0
R2	0,0	1,1

Figure 3.1

[2] The requirement of a *strict* Nash equilibrium is not incidental: this permits to rule out mixed-strategy equilibria as possible candidates to the role of solutions of co-ordination problems. In fact, a mixed strategy equilibrium would arise relevant problems in the formation of reciprocal expectations.

Predominant coincidence of interest can thus be said to obtain when the interaction is 'sufficiently' close to symmetrical games; that is to say, the game has mutually beneficial Nash equilibria but is not necessarily symmetrical. An instance of a game with predominant, but not perfect, coincidence of interests is the 'Battle of the Sexes': in fact, agents have conflicting interests on which of the equilibria is to be selected. At the other side of the spectrum, instead, there lie games of pure conflict, i.e. zero-sum games (Schelling (1960: 83-118, 291-303)).

Summing up, on Lewis's account a co-ordination problem arises when *at least two strict mutually beneficial Nash equilibria exist in a game.* The problem lies in that, in spite of the presence of coincidence of interests, the presence of more than one of such mutually beneficial equilibria requires that some additional piece of information are added to the game in order to elicit co-ordination between them. This leads us to the analysis of the concept of expectation, upon which a more complete definition of mutually beneficial convention will be put forward in section 3.1.3.

3.1.2 Expectations

3.1.2.A Systems of Expectations

Many authors stress how the main character of a convention lies in its function of making the agents' expectations self-fulfilling[3]. In fact, it is evident that each rational agent will strive to form an *expectation* about others' behaviour in order to choose her optimal action. Such expectations will be an effective means to 'solve' the co-ordination problem only if they are *reciprocal* - that is, they are based on a conjecture about others' behaviour- and *concordant* – namely, they must lead to the same co-ordination equilibrium. If these two properties are satisfied then a *system of expectations* can be set up (Lewis (1969: 25)).

But what is the domain of one's expectations and how can they be constructed? The two questions are related. Three pieces of information are required to form an individual's expectations:

---

[3] In Hayek (1973), this is indeed the fundamental notion in his concept of order of a society. For a critical account, see Sacconi (1986), and Bicchieri (1990: 840).

a) others' preferences about the outcomes;
b) others' degree of rationality[4];
c) what one believes about the matters of fact that determine the likely effects of others' alternative actions (Lewis (1969: 27)).

Undoubtedly the last element is the most controversial: in fact, it suggests that each agent will attempt to replicate others' reasoning to predict what their actions will be. This will generate a *chain* of mutual expectations, reaching an *order* of infinite level. Notice that this infinite-long chain of expectations exists only on a mere *logical* ground. They may be called a *support* to expectations, as only adding ancillary hypothesis of rationality to them proper expectations can be formed[5].

3.1.2.B Devices to Generate Expectations and Common Knowledge

The analysis conducted so far only explains the *formal* structure of one's expectations, but as yet it is silent on their *substantive* content. Clearly, some elements not directly linked with the structure of the game are needed in order to shape one's belief on others' behaviour. In particular, Lewis distinguishes three elements that are able to confer a substantive content to expectations: these are given by *agreements*, *precedent* and *salience*. The first two instances imply that some action has already taken place *before* the occurrence of the interaction described by the game, i.e. some communication or even an instance of a successful co-ordination. The latter element is probably the most interesting in that it requires no action being taken place earlier, but the sharing of some common cultural or cognitive trait by the agent that makes one outcome "*standing out from the rest by its uniqueness in some conspicuous respect*" (Lewis (1969: 35); also, Schelling (1960)). A large amount of experimental evidence has been gathered regarding this notion, even though a formal treatment still seems to lie outside the theorists' agenda (for an attempt in this sense, see Binmore (2002)).

---

[4] Lewis allows for the fact that agents may not have a "full" rationality, or that they are more or less likely to commit mistakes. As a matter of fact, a modicum level of rationality in the agents is enough to generate mutual expectations (Lewis, 1969: p. 27).

[5] Sugden (1990) stresses how the attempt to replicate others' reasoning is problematic in a context in which there exist more than one equilibrium, and criticizes theories of bargaining for their reliance on some assumptions that *covertly* make such an operation possible. See in particular the crucial role of the "conditions for strategically rational choice" in Gauthier's argument (Gauthier (1986: 61)).

What agreement, precedent and salience all have in common is that they are means to generate concordant expectations. Formally, Lewis defines a state of affair A as a *basis for common knowledge* if it meets the three following conditions, which for simplicity will be stated for a two-person interaction:

(1) You and I have reason to believe that A holds;

(2) A indicates to both of us that you and I have reason to believe that A holds;

(3) A indicates to both of us that X, where X could be any property of the interaction in which we are involved, and in particular the fact that one of us will follow a certain action.

A could be either the content of our agreement stipulated before the game, or the characteristic that makes an outcome salient, or the fact that a precedence of successful co-ordination has occurred. In all these cases A *indicates* to us a basis of knowledge to extend our expectation about others' behaviour.

If we iterate the application of each condition to itself and to the others, we are able to generate the infinite-long chain of implication about others' behaviour, which forms the support necessary to build our expectations. For instance, (2) applied to (3) implies:

(4) A indicates to both of us that each of us has reason to believe that you follow the action X;

And (2) applied to (4) implies:

(5) A indicates to both of us that each of us has reason to believe that the other has reason to believe that you follow the action X;

Formally, we shall say that it is common knowledge in a population P that X if there exists a state of affairs A such that conditions (1) to (3) are satisfied.

If these statements are accompanied by ancillary premises about our rationality, then we are allowed to substitute the clause "has reason to believe" with the clause "expects". In fact, while the chain of logical implications does not require any hypothesis about our rationality -since they include the "neutral" clause *"has reason to believe"*- our expectations do need them.

3.1.3 A Definition of Mutually Beneficial Convention

On the basis of the analysis set out above, Lewis summarises the essential features of a convention in the following elements (Lewis (1969: 69)):

a) each agent involved in an instance of S prefers to conform to R conditionally upon conformity by the other agents involved in S;

b) all agents involved have approximately the same preferences regarding combinations of their actions, so that S is a situation in which coincidence of interests is predominant;

Taking these two conditions together, we obtain the condition of mutual benefit: not only do I prefer to conform to R when all the others do, but also I prefer that each of the others conform if all except that agent are conforming.

c) there is (at least) a second possible regularity R' in S which meets the same conditions we are imposing to R.

Lewis stresses that R' must share with R all the characteristic necessary to make it a convention; in order words, R' *could* have become a convention if only, for some reason, the agents had started off co-ordinating their behaviour on R' instead of R. Not only must the alternative R' share the formal requirements in order to be a co-ordination equilibrium, but also it must be similar to R for its *substantive* content. For instance, if R' gives each agent a payoff remarkably lower than R, then R' should not be considered a possible alternative to R (Lewis (1969: 73)).

On the grounds of these last remarks, we can put forward a refined definition of convention:

A regularity R *in the behaviour of members of a population P when they are agents in a recurrent situation S is a convention if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P,*

(1) *everyone conforms to* R*;*

(2) *everyone expects everyone else to conform to* R*;*

(3) *everyone has approximately the same preferences regarding all possible combinations of actions;*

(4) *everyone prefers that everyone conform to* R*, on condition that at least all but one conform to* R*;*

(5) *everyone would prefer that everyone conform to* R', *on condition that at least all but one conform to* R',

where R' *is some possible regularity in the behaviour of members of* P *in* S, *such that no one in any instance of* S *among members of* P *could conform both to* R' *and to* R.

Recalling the distinction set out in the introduction, and furthering some arguments that will be better clarified in section 3.4.1, we can immediately notice how in the case of mutually beneficial conventions there is no opposition between self-regarding and other-regarding motives to action, since social rationality and individual rationality point in the same direction.

3.2 Individually Beneficial Conventions

As illustrated in the foregoing section, the basic features of a mutually beneficial convention are the presence of predominant coincidence of interests between the agents and the existence of more than one Nash equilibrium with comparable rewards for the participants. However, Robert Sugden argues that one can talk of a convention even when the former element is missing. This view enables him to classify various kinds of regularities of behaviour as conventions, covering almost the whole range of interactions that characterise a society. What all of these rules have in common is their 'conventional' character, in that they are all arbitrary patterns of behaviour that emerge after some process of 'selection' between the rules themselves has taken place. In fact, the final goal of Sugden's work is to provide a comprehensive theory of the sense of morality consistent with Hume's theory of moral conventionalism. There are two aspects in his notion of convention, which also corresponds to two different stages of his work, which in turn emphasise the equilibrium character of the equilibrium and the dynamic process that leads to it. I will break down these two aspects and analyse each of them separately. I will deal with the static characteristic in the present section, whilst leaving the dynamical one to section 3.6.

3.2.1 The Notion of Evolutionarily Stable Equilibrium

In Sugden (1986: 32) a convention is defined as any Evolutionarily Stable Equilibrium (ESS) in a game that contains two or more of such equilibria. The

concept of ESS carries with it a dynamical aspect that will be further explored in section 3.6.1, but in the present section I focus on its 'static' character. Simply stated, the idea of an ESS is that it is an equilibrium is able to 'repel invaders'. That is, suppose that the population of agents is behaving according to some profile σ and then a small proportion ε of 'mutants' start playing a different strategy σ'. ESS asks that, in order for σ to be an equilibrium, it must be robust to such a type of 'invasion'. That is, the existing population gets a higher payoff against the resulting mixture *(1-*$\varepsilon)\sigma + \varepsilon\sigma'$ than the mutants do. In formal terms, in order for σ to be an ESS, we require that $u(\sigma,(1-\varepsilon)\sigma+\varepsilon\sigma') > u(\sigma',(1-\varepsilon)\sigma+\varepsilon\sigma')$ for any other $\sigma' \neq \sigma, \sigma' \in \Sigma$ and for all sufficiently small ε.

The main implication of this concept for our purposes is that it can be shown that every ESS is a strict Nash equilibrium in the stage game (see Fudenberg and Levine (2000: 59)). This implies that, with respect to Lewis's definition of co-ordination equilibria, the requirement of individual rationality – i.e. the Nash equilibrium condition- still holds, while that of a mutually beneficial equilibrium is dropped. This makes it possible to define an 'individually beneficial' Nash equilibrium, and to call 'individually beneficial conventions' the associated type of interaction. In what follows I will give some examples drawn from Sugden of this kind of convention.

3.2.2 Conventions of Property

The hawk-dove game can be interpreted as a simple model depicting the interaction about the allocation of property rights in absence of any pre-constituted authority or agreement. Its payoffs matrix is reported in Figure 3.2:

	Dove	Hawk
Dove	1,1	0,2
Hawk	2,0	-2,-2

Figure 3.2

The main assumption in such a situation is that when people *desire the same thing, which nevertheless they cannot both enjoy*, they may either incur in a fight, generating a mutually destructive outcome, when both claim the good (Hawk-Hawk), or in a peaceful sharing of the good (Dove, Dove), or finally in the attribution of the thing to one of the two, when only an agent claims the good and the other renounces. The ESS of the game are given by the pair of outcomes in which one agent plays Hawk and the other Dove. Therefore, the convention that will emerge will be one in which the property of the good is assigned to one of the two agents. Notice that either equilibrium is not mutually beneficial: the agent who is playing Dove would be better off if the other played Dove as well instead of Hawk, which is the strategy prescribed by the convention. As we shall see, this has some consequences on the cognitive structure needed to sustain a convention.

New insights can also be added into the game, as will be shown in section 3.6.3, by relying on the evolutionary parable. In particular, the mixed strategy of the game, which leads to a sub-optimal outcome in terms of payoffs, can be associated with the situation of stalemate in a Hobbesian state of nature. However, it will be shown how a convention can spontaneously emerge from that point with no need of any outside authority, as is the case in Hobbes' theory. Hume's claim that property rights are the result of a process of convergence to a convention when individuals interact in a state of nature is therefore given a formal treatment through this model.

3.2.3 Conventions of reciprocity

The typical example of a reciprocity game is given by the well-known Prisoner's Dilemma (PD from here on). Clearly, in the static game there is no possibility for a convention to emerge, given the existence of a unique Nash equilibrium. However, many commentators have argued for the possibility of defining a convention even in this case (Sugden (1986: ch. 6); Hardin (1988)). If we consider the *super-game* made up by the repetition over a possibly infinite period of the one-shot PD, with a positive probability of the super-game suddenly ending after a finite number of periods, an infinitely large number of possible ESS exist. Picking up for simplicity just two of

these, that is the well-known *tit-for-tat* [6]and the strategy prescribing to defect at every game - the so called *nasty strategy*-, we end up with the following matrix of payoffs gained in the super-game:

	Tit-for-tat	Nasty
Tit-for-tat	10,10	-1,3
Nasty	3,-1	0,0

Figure 3.3

This is a coordination game with two ESS, so that both tit-for-tat and the nasty strategy satisfy the requirements necessary to define an individually beneficial convention. Sugden is therefore confident that even in this relevant class of interactions the rule that emerges in the society is after all a matter of convention.

This account may be criticized on the grounds that the couple of ESS here considered falls short of one of the constitutive properties of a convention, namely their similarity (see section 3.1.3). In fact, the equilibrium given by (T,T) will give a remarkably higher reward to the agents if compared with the equilibrium (N,N), thus casting some doubts over the possibility of considering each equilibrium as a proper alternative to the other.

A way of escaping this problem may be provided by considering that there exist a large number of tit-for-tat-like strategies, depending on the number of periods of punishment prescribed after the defection of the opponent. If we allow for the possibility that agents make mistakes in their actions, so that, say, they can defect with a small probability, even though their strategy requires them to Cooperate, then a typical coordination problem with 'proper' alternatives can be set up in this context. For when an agent makes a mistake and fails to Cooperate, then an infinite series of retaliations occurs if the agents had not previously coordinated on the same 'type' of tit-for-tat. Conversely, when the agents abide by a tit-for-tat with the same length of punishment, Cooperation can be restated swiftly after some interactions. Although this setting 'formally' makes up a typical coordination problem with comparable

[6] For the appraisal of the properties of tit-for-tat in reciprocity games, and its comparison to other strategies, see Axelrod (1984).

alternatives, the fact that the difference between these can only be appreciated *off-equilibrium* still makes it at least doubtful that such a situation is characterised by a 'conventional' character.

3.3 OTHER-REGARDING CONVENTIONS

In the conventions analysed so far, the equilibrium strategies were upheld by some form of self-interest, which was beneficial either individually or mutually, i.e. for all the players involved. In the present section, instead, I introduce regularities of behaviour that, despite being detrimental in terms of self-interest, can nonetheless emerge as 'equilibria' of games. I call these *other regarding* conventions to emphasise the concern that at least some players must have for some non-self-interested aspects in order for the convention to be sustained. The strategy I follow to define this third category of convention is to rely on the work of some authors, notably Lewis (1969) and Pettit (1990), whose definitions refers to the wider category of *norms*. That is, this notion includes as a special case the previously defined mutually and individually beneficial conventions. Hence, the specific character of other-regarding conventions can emerge through the contrast of this wider definition with the previous ones, as the following analysis will show.

The main character of a norm is highlighted by David Lewis: *a regularity in behaviour to which we believe one ought to conform* (Lewis (1969: 97)). It is then apparent that the relevance of a norm lies in the aspect of *obligation* that an agent must experience in complying with it. Philip Pettit elaborates on this point arguing that there are three constitutive elements in order to classify a regularity of behaviour as a norm:

A regularity R *in the behaviour of members of a population P, when they are agents in a recurrent situation S, is a norm if and only if, in any instance of S among members of P,*

(1) nearly everyone conforms to R*;*

(2) nearly everyone approves of nearly anyone else's conforming and disapproves of nearly anyone else's deviating[7]*;*

[7] It should be pointed out that between the condition of approval of compliance and disapproval of deviance, the crucial one is the second. In fact, if we disapprove of someone's not doing an action and do not disapprove of her doing the action, this is tantamount to approving of the action. Conversely, we can approve of someone's performing an action and not disapproving of her not performing it,

(3) the fact that nearly everyone approves and disapproves on this pattern helps to ensure that nearly everyone else conforms. (Pettit (1990: 731))

Such a definition is clearly moulded on Lewis's one. What appear to be the most relevant difference is that now there is no requirement on individual self-interest; in fact, the key aspect is that of the sentiment of approval and disapproval to the norm that substitutes for the conditions on individual preferences and reciprocal expectations in the previous definitions. To be sure, if the norm satisfies individual self-interest, then approval in case of conformity and disapproval in case of deviance are expected, especially in situations of coincidence of interests between individuals. However, this is not necessary: the 'sense of obligation' associated with the norm may provide sufficient motivational force on agents to elicit compliance with the norm. This feature helps to sustain regularities of behaviour in which, in the language introduced in Chapter 1, other-regarding motivations offset self-interested ones to uphold the convention.

One aspect that will become crucial throughout the first part of the thesis is the content of such a sense of obligation: does this need to rest on some justification that is somehow 'external' to the interaction, such as, in particular, some ideas that the action is 'collectively rational', though individually onerous? Or are the 'internal' elements to the interaction, such as the sense of obligation elicited by the expectations of other agents, sufficient to elicit compliance? The thread of the analysis in this and the next chapter will be given by the attempt to answer this question.

Many authors seems to take as granted, that, in most cases, the norm will fulfil some notion of public interest: this will provide the agents of a community with a 'valid' other-regarding reason to comply with the norm. The easiest case is the one in which an outcome is Pareto-superior to all others in an interaction. Typical is the case of a PD-like situations. For instance, Pettit deals with patterns of behaviour satisfying what he calls the *interaction assumption: "among the options available to any agent in the sort of situation involved nearly everyone is better off if everyone else takes one particular option than if everyone else rejects it: the option in question is, in that sense, a collectively beneficial one"* (Pettit

---

but this situation falls in category of super-erogatory virtues, which should not be considered norms (Pettit, p. 730).

(1990: 743)). Notice that such a condition does not require that the action prescribed by the norm uphold *anyone*'s self-interest, but that it is reciprocally beneficial for everyone, thus comprising Pareto dominant allocations. This argument may provide an explanation for the emergence of cooperation even in static PD. Likewise, Bicchieri (1990) explicitly focuses on repeated PD-dilemmas where agents can take on some predetermined types of strategies.

However, other authors stress how other-regarding conventions may not lead to collectively beneficial outcomes, at least in the clear-cut sense that Pareto-dominancy implies (Sugden (1998a)). Such types of interaction could be classified in the general category of 'rules of courtesy', in that some parties are required to yield to the benefit of some other parties. For example, the custom to offer the seat on a bus to an elderly person can be justified by the desire to satisfy her needs, but such a behaviour causes a disadvantage to the person who renounces to his seat, so that the action change is not Pareto-improving for either party. Also, the non-written rules of traffic on the road often prescribe a driver to give way to other cars although the 'formal' rules allow him not to. For instance, this happens when some drivers are being stuck in a long queue at a crossroads whereas one lucky driver has found his way clear throughout; in this case, the latter will feel in some sense obliged to give way to some of the other cars. The norm seemingly aims to make up for the bad luck of some drivers by asking an action contrary to his self-interest to the fortunate driver.

In all these situations, a sense of public interest, although as not unambiguous as in the Pareto-dominance sense, may perhaps be found by looking at that particular phenomenon called 'multilateral reciprocity' (Hollis (1998)). This is the case when an individual gives in to an agent in the current situation, 'because' he envisages to be in the position to benefit in the future with respect to a third party, who is not currently involved in the interaction but is part of the 'community', thus he is expected to conform to the same rules. This type of reasoning may even extend its scope across generations, as the example of giving one's seat on the bus shows.

As stressed by several authors (see Elster (1990: 885) and Frank (1988)), though, the exercise of finding some notion of 'public' interest behind the consequences of a norm is seemingly impossible in some instances. Typical in this respect are norms of

revenge, which force individual to undertake potentially self-destructive actions for the purpose of what, at best, can be seen as a private form of reward deriving from the acquisition of social status. In fact, in this case, by retaliating a person seems to uphold the interests of his family, or clan, or to boost his reputation, although at a very high risk, but it is hard to see how this could have a 'public' beneficial impact. To be sure, one could think at norms of revenge as a form of 'endogenous enforcement of social rules of conduct', cruel as these may be, so that they would indeed carry with them some idea of public interest. But this is arguable, and, to say the least, not the most efficient way to reach the goal.

Nevertheless, despite the prevailing 'private' character that these norms have – namely, their being justified by a private, as opposed to a social, or public, reward – these rules bear the typical traits of norms in that they are somehow expected by other members of the community, and the compliance with them or the failing to comply may indeed trigger approval or disapproval. In fact, this is a key aspect of social norms: a norm has to be *socially enforced*, i.e. publicly sustained by people not directly involved in the interaction. On Lewis's account, a failure to conform by some agent tends to evoke unfavourable responses from *all* others agents, even those not directly involved in the instance (Lewis (1969: 99)). Going back to Hume, two reasons could be found to explain this. The first is 'immediate': we typically experience a 'natural' sentiment of *sympathy* toward others as they go through a certain situation, which is generally the stronger the closer the affective relationship with them. The second is only 'mediated', and refers to that kind of 'multilateral reciprocity' that I mentioned earlier: we could think that in the near future we could be involved in the same kind of situation of which now we are only bystanders, thus having a reason to perpetuate the norm as much as we can. Therefore, we may be led to condemn someone breaching a norm for the fear to meet him in our next interactions, or for the knowledge that the imitation on a wide scale of such a deviant behaviour by other people could end up to be detrimental for my interests.

3.4 The Problem of Compliance with Norms and the Question of their Moral Content

So far we have only surfaced of the reasons to comply with them, especially in the case of norms sustained principally because of other-regarding reasons. Strictly connected to this question is that regarding the moral content that can be attributed to those patterns of behaviour. This is the subject of the present section.

3.4.1 Presumptive Reasons for Conformity to Norms: Mutually Beneficial Conventions

The point of departure in accounting for the reasons to comply with conventions and norms is the same, and consists of the so-called presumptive reasons to conform to a convention. Mutually beneficial Nash equilibria are sustained firstly because of self-interest. To be sure, this gives a straightforward reason to comply with them. But the contemporary presence of the element of the reciprocal benefit in the pursuing of one's self interest also attributes a specific normative and moral character to these regularities.

Let us investigate in further detail this concept. From Lewis's definition of convention we can derive the following implications:

(1) Most other members of P involved with me in situation S will conform to R;

(2) I prefer that, if other members of P involved with me in S will conform, then I also conform;

(3) Most other members of P involved with me in S *expect, with reason*, that I will conform;

(4) Most other members of P involved with me in S *prefer* that, if most of them conform, I will conform;

(5) I have reason to believe that (1)-(4) hold.

Therefore, the concept of mutual benefit underlying a convention is crucial in order to derive such implications. Finally, the clause (5), applied to the other four sentences gives:

(6) I have reason to believe that my conforming would answer to my own preferences;

(7) I have reason to believe that my conforming would answer to the preferences of most other members of P involved with me in S; and that they have reason to expect me to conform.

(6) and (7) are what we may call presumptive reasons why I ought to conform: *for we do presume, other things being equal, that one ought to do what answers his own preferences. And we presume, other things being equal, that one ought to do what answers to others' preferences, especially when they may reasonably expect one to do so* (Lewis, p. 98).

Of the *two* reasons, the first is related to one's own preferences, the second is connected to others'. While the former is a simple restatement of a self-interested motivation, the latter introduces an other-regarding motive to action: I ought to do what is in the interest of others.

Although the adverb "especially" seems to imply that in some occasions it might be true that one follows a certain behaviour only having in mind her other-regarding motives to action, thus satisfying others' preferences without considering her own, Lewis seems to consider this occurrence a rather exceptional event. In fact, in the following passage he clearly claims that the two presumptive reasons must both be present in order to support a convention: *"for any action conforming to any convention, we would recognise these two (probable and presumptive) reasons why it ought to be done"* (Lewis (1969: 98)). In this sentence, the adverb "reasonably" refers to the fact that by following the convention I am responding to my preferences, as I will clarify later on. The others have reason to believe that I will conform as long as they know that it is in my interest to conform.

This point is further clarified by Lewis when he deals with socially enforced norms. On Lewis's account, the sentiment of disapproval I invoke in people who are part of my society comprises both a feeling of resentment for not having answered others' preferences, and a feeling of surprise to have acted contrary to my own preferences. *"So if they see me fail to conform, not only have I gone against their expectations; they will be probably be in a position to infer that I have knowingly acted contrary to my own preferences and their*

reasonable expectations. They will be surprised, and they will tend to explain my conduct discreditably (Lewis (1969: 99)).

It is for the presence of these two kinds of "ought" derived from the pair of presumptive reasons for conformity that the normative character of this type of convention becomes apparent. It is indeed straightforward to show that the three conditions set out in section 3.4 in order to define an other-regarding convention are satisfied: of course they commence a regularity of behaviour (condition 1), which elicit commendation on conformity and censure on deviance (condition 2) from all the members of the community, and every agent can understand that these sentiments elicited in members of the community are reasonable, thus strengthening the compliance with the norm (condition 3).

Furthermore, we can expand on this argument and recognise an underlying moral idea in this kind of patterns of behaviour. Sugden calls it the *principle of co-operation.* He claims: *the moral rules that grow up around conventions are likely to be instances of the same principle: Let* R *be any strategy that could be chosen in a game that is played repeatedly in some community. Let this strategy be such that if any individual follows* R, *it is in his interest that his opponent should do so too. Then each individual has a moral obligation to follow* R, *provided that everyone else does the same* (Sugden, 1986, p. 172). In this case, it is the very fact that agents form *reasonable* expectations about others' behaviour, where reasonable means conforming to self-interested and reciprocally beneficial actions, which attaches a specific *moral* obligation to those actions.

3.4.2 Presumptive Reasons to Conform: Individually Beneficial and Other regarding Conventions

The two presumptive reasons are both present in the narrower concept of conventions. As far as the equilibrium is mutually beneficial, I both have an interest in following the rule and others have an interest that I follow the rule, thus giving me an additional motive to follow the rule. The problem with individually beneficial conventions is that they are not always mutually beneficial equilibria. Such a problem is even made worse in other regarding conventions, where, as pointed out earlier, self-interest may be lacking as a motivational source. On Sugden's account, however, a

moral reason to abide by individually beneficial conventions and other regarding conventions can be grounded on the concept of normative expectations and in particular on the possibility of considering the two presumptive reasons set out above *independently* from each other (Sugden (1998a); (2000a)). In fact, the knowledge of others' expectations takes on the status of a moral commitment in virtue of the idea of normative expectations.

Sugden introduces his analysis of normative expectations explicitly recalling Lewis's argument. He then argues that, between the two presumptive reasons for conformity, the second is the crucial one (Sugden (1998a: 9)). To support this idea, Sugden argues that the fact that others have a *reasonable* expectation on me following a certain behaviour is derived from the fact that once a convention has established, the precedent acts as a powerful force in order to shape individuals' expectations about others' behaviour. He defines such a kind of expectations as *well-grounded empirical expectations*, thus emphasising the importance of past experience in the process of anticipating others' behaviour.

On the grounds of such a reading of Lewis's presumptive reasons, it is easy to carry over the same argument to the most general case of regularities of behaviour that are *not* mutually beneficial. Sugden claims that *Lewis's* [second] *presumptive reason makes no explicit reference to conventions. It is stated as an entirely general principle, referring only to actions, preferences and reasonable expectations* (Sugden (1998a: 11)). If this was true, then we may identify other-regarding motives to obey to a regularity of behaviour even if this is not a mutually beneficial Nash equilibrium and, more noticeably, even if this is not a Nash equilibrium at all. The point is that whenever a rule is well established, others people have an 'empirically-based' *reason* to think that I will adhere to it, even if this is contrary to my interests. It is the force of precedent that allows them to form this expectation. It is sufficient that the second presumptive reason has been formed in other to prompt the agents to follow the prescribed behaviour: the resentment hypothesis will make them feel in some sense *obliged* to adhere to it, thus providing it with a normative and moral content.

3.5 Some Critical Remarks on the Concept of Normative Expectation

3.5.1 Reasonable Expectations

On Sugden's account, for normative expectations to act as a guide to action it is sufficient that a regularity of behaviour has established. Whenever agents know that, then they will form the expectation that agents behave in a way coherent to such a regularity. Notice the different use of the adjective 'reasonable': while in the case of mutually beneficial conventions this both implied a self-regarding and an other-regarding motivation (section 3.4.1), in the case of more general norms one can find 'reasonable' whatever type of behaviour is enrooted in the habits of a community and has acquired sufficient regularity, notwithstanding the relationship to someone's self-interest (section 3.4.2).

For example, if, for whatever reason, it happens that the challengers for the possession of an object give in in many repetitions of the game, than in the next instance of the game the possessor has a *reason* to expect the challenger to surrender as well. As the challenger shares the same information of the possessor, he knows that the possessor is likely to form that expectation. As the expectation has a normative content, he will feel urged to act with remission, thus confirming the expectation of the agent.

In this and the in next section I would like to advance two critical remarks to such a use of the concept of empirically grounded expectations. Both of them are related to the intrinsic stability of a norm sustained on expectations: the first deals with the possibility of grounding on precedent a basis of common knowledge for reciprocal expectations, the second explores the internal logical structure of conformative behaviour.

3.5.2 When is an Expectation Reasonable? Empirical and Causal Expectations

The argument in this section is related to the use of the term 'reasonable' that allows Sugden to introduce the concept of well-grounded empirical expectations. As

his starting point is Lewis's work, it seems appropriate to come back to his work. Throughout his book, Lewis uses the expression "to have reason to X", where X can be either the verb "to desire", or "to expect", or "to believe", as a part of a peculiar inference, which has the following structure:

(A) If I *desire* that I perform X on condition that you perform Y and I *expect* that you perform Y, then I have reason to perform X.

Further, translating this inference to a higher order,

(B) If I *expect* that you *desire* that you perform Y on condition that I will perform X and I *expect* that you *expect* that I will perform X, then I have reason to expect that you have reason to desire that you perform Y.

Therefore, it is apparent that Lewis uses the term 'reason' when drawing inferences involving *both* preferences and expectations. Consequently, one can infer that when Lewis claims that people can *reasonably* expect that I will follow a certain behaviour, not only is this grounded on the fact that they have observed me conforming in the past, but also because it is in my interest to do so. As Lewis deals with mutually beneficial Nash equilibria, the most powerful reason that people have in expecting that I will follow the convention is given by the fact that I would be worse off if I breached the rule. This is the reason why they would be *surprised* observing me disobeying to it (section 6.1.1). Clearly, the surprise does not consist of the resentment for me not having lived up to their expectations, but because of the irrationality of my action.

This argument may be restated by saying that there exist two types of expectations, depending on the kind of information that is common knowledge amongst the agents. The first are *causal expectations*, and are derived from a series of inferences related to the *preferences* of the agents. On such grounds, I might say that I have a reasonable expectation of X, because I know that X is convenient for you. The second are what Sugden qualifies as *empirical-grounded* expectations, and are drawn from the precedent occurrences of the interactions without necessarily referring to the preferences of the agents involved.

Hence, it seems that we can conclude that on Lewis's account both kind of expectations, causal and empirical, must be present in order to make up a firm *reasonable* expectation. We can infer that it *might* be the case that people form their

expectations on the ground of a precedent, thus forming empirical expectations only, but this would give a much less stable basis for the concordance of expectations. This remark clearly does not undermine Sugden's point, but stresses how the claim that *all* of Lewis's accounts of reasonable expectations are grounded on an empirical basis seems to miss some important aspects of his accounts of norms.

3.5.3 Conformative Behaviour and Normative Expectations

In this section, I will try to deepen the analysis of normative expectations, investigating to what extent this idea could be relied upon in generating moral support to an action. This will supply the basis to put forward a second criticism to this concept. What I wish to show is that Sugden's idea of normative expectations comes very close to what Lewis treated as *conformative behaviour* (Lewis(1969: 107-118)). Hence, I shall try to restate Sugden's argument in terms of this concept, in order to make clear its internal cognitive structure.

A typical choice to adhere to a rule standing on a conformative behaviour can be restated by means of the following inference:

First premise: I desire that I conform on condition that you expect that I will conform;

Second premise: I expect that the existence of a rule entitles you to expect that I will conform;

Conclusion: I conform

In my view, this inference seems suitable to describe Sugden's argument. The first premise is derived from the second presumptive reason for conforming, whereas the second is an inference from the observation of past occurrence of the rule. As usual, the motivation to act is derived from a preference and an expectation. However, in this peculiar case, both one's preferences and expectations depend upon other's expectation.

The problem of this inference is that it explicitly relies on the rule itself, creating a circularity in the definition (the existence of the rule shapes my expectations, but actually it is a system of concordant mutual expectation which should give rise to a

rule). However, according to Sugden, we actually do not need the concept of rule in this inference, but only the occurrence of a precedent. In other words, the precedent acts as a basis for common knowledge about our expectations. Therefore, we can restate the previous inference eliminating any explicit reference to a rule:

First premise: I desire that I conform on condition that you expect that I will conform;

Second premise: I expect that you expect that I will conform;

Conclusion: I conform

In this syllogism, my n-th order expectations are generated by means of a precedent that, acting as a basis for common knowledge, allows us to derive expectations using higher order expectations:

First premise: I expect that you expect that I desire that I conform on condition that you expect that I will conform;

Second premise: I expect that you expect that I expect that you expect that I will conform;

Conclusion: I expect that you expect that I will conform

As usual, to generate an expectation of the n-th order about one's behaviour, two types of further expectations are needed: a (n+1)-th order expectation about one's behaviour and a n-th order expectation about one's preferences.

Now, we can notice that the precedent only helps to generate expectations of highest order for 'proper' expectations about other's behaviour, i.e. the second premises of the former inference. As I observed you conforming yesterday, the day before yesterday, and so on, and you saw me conforming as well, then we have reason to believe that each of us will conform tomorrow. However, it would be wrong to say that the same information serves as a basis of common knowledge for desires as well, especially when these desires are conceived to be dependent on others' expectations about one's behaviour. Even if you saw me conforming in the past, you cannot infer that I *desired* to do it, and above all, that in doing so I took your expectations into account. This is to say that empirical and causal expectations must rely upon a different basis of common knowledge, or at least that the informative content to derive causal expectations is much wider.

What I wish to stress with this argument is that people *may* actually draw inferences from the past experience in order to learn about others' preferences: it is clearly sensible that, if I have always seen you conforming to a rule, I may infer that this was your *real* desire. But the information provided is not adequate to form a basis for common knowledge. Resting upon Lewis's definition, what we need is some element - or state of affairs - that *indicates* to all of us that I desire to conform on condition that you expect me to do so. In my opinion, the precedent cannot provide this unambiguous state of affair, especially considering that these preferences are supposed to depend on what is your expectation about my behaviour.

In mutually beneficial conventions, we do not encounter this problem since it is not possible to form an expectation about a behaviour contrary to one's self-interest. But for other regarding conventions sustained on other-regarding motivations only, this is not necessarily guaranteed. Conformative behaviour shows a considerable lack of stability when it is not accompanied by self-interested motivation. If we do not take into account the 'objective' consequences of everyone's expectations on everyone else's outcome, the interplay of expectations may well turn out to be completely void of "intrinsic" significance, and it is hard to believe that rational individuals will constantly adhere to the related regularities of behaviour. As Lewis puts it, in conformative behaviour, *I would be the only agent in the situation; the others would be involved merely as supposed holders of expectations about me. [...] An important fact about the intended sort of conformative behaviour is left out: namely that I want to conform to your expectation <u>because of</u> the way I expect you to act on your expectations. If I thought you would not act on your expectations, I would concern myself with how you would act, not with what you expect. When this fact is left out the story, our understanding of the phenomenon is badly distorted* (Lewis, pp. 114-116).

The most relevant problem with Sugden's theory is that expectations seem to be a set of stable parameters easily recognisable by each agents. Take the case of a repeated Prisoner's Dilemma, and suppose that the current rule prescribes that one agent Defects and the other Co-operates. Formally, normative expectations would suffice to sustain such an outcome, since the gain that the co-operator would get if he Defected could be outweighed by the resentment for not having lived up to others' expectations within her overall interests function. But what about the way their expectations are

formed? The Defector knows that the rule imposes a heavy cost on C, thus she may fear that at some later stages of the interaction C will change his mind about serving D's own interests. Her expectation about C's conformity cannot be very strong, if she takes into account the opponent's direct interests. Further, if C is sufficiently rational, she will anticipate D's doubts about her behaviour. Therefore, she may perceive that D's expectations about her are not so strong as she believed in the past. When forming her second order expectations, she may take this into account, and believe that her expectation that D expects her to conform is not so firm. Then D should allow for C's perplexity when forming his third order expectations, and so on. The point is that nearly all of the outcomes of the game may be considered equilibria if sustained by an opportune set of expectations. And to point to normative expectations in order to elicit compliance to a convention is tantamount to saying that in a society dominated by slavery slaves conform to the rule because of the sense of resentment they would elicit in their masters. Expectations based on empirical reasons but not grounded on individual interests would fail to be a basis for common knowledge: with rational agents the expectations would not extend much far beyond the first levels of mutual expectations, thus undermining the stability of the rule.

Claiming that expectations make up all the normative content of conventions is in my view implausible: surely they help to strengthen the commitment to a norm whenever this is established, but they cannot *per se* exhausts the set of motivations underlying the norm. In Bicchieri's words (1990: 839), *to say that one conforms because of the negative sanctions involved in nonconformity does not distinguish norm-abiding behaviour from an obsession, in which one feels an inner constraint to repeat the same action in order to quiet some "bad" thought*; or, to put it in other words, actions complying with norms may be seen as a categorical imperative[8]; but then the reductionist endeavour of grounding social institutions on individual deliberation may be put in doubt.

I instead believe that some additional feature is required in order to account for the whole cognitive system beneath a set of concordant mutual expectations and the sense of justice people experience, and this further element should come from some idea of public interest that people perceive in a rule, of which the concern for others' *individual*

interests involved in the interaction, transmitted through others' expectations, is a constitutive part.

This account echoes Hume's considerations on the subject:

"Conventions turn out to be a general sense of common interest; which sense all the members of the society express to another, and which induces them to regulate their conduct by certain rules. I observe that it will be in my interest to leave another in the possession of his goods, provided he will act in the same manner with regard to me. When this common sense of interest is mutually expressed and is known to both, it produces a suitable resolution and behaviour. And this may properly be called a convention or agreement betwixt us, though with the interposition of a promise; since the actions of each of us have a reference to those of the other, and are performed upon the supposition that something is to be performed on the other part" (Hume (1740: III.ii.2)).

However, this approach may be bound to fail as well. Bichieri associates the attempt to justify a social norm on the grounds of its *collective* rationality with a *post hoc ergo propter hoc* fallacy. This would in fact be the case if, say, we said that a norm was established because of its efficiency in pursuing a certain end, like for instance social welfare, since the mere presence of a social norm does not justify inferring that it is there to accomplish a social function, and indeed in many cases the contrary may be held (Bicchieri (1990: 838)).

However, I believe that the argument put forward in this section, which will be further expanded in the next Chapter, is safe from this type of criticism. In fact, my claim is that the perception that public interest is being furthered through an agent's action may give the agent the 'causal reason' to abide by it, and also give all the agents a sufficiently sound basis on which to form expectations. For public interest may replace individual desires in the construction of the chain of expectations reconstructed earlier on in this section, because both are criteria of assessment 'independent' from the current interaction. Normative expectations could not deliver on this aspect, as they would be determined by the interaction itself, whereas the system of expectations to be concordant needs an 'externally' determined set of preferences of the individual on outcomes. In other words, they are immune from the criticism of 'circularity' implicit in the normative expectations argument. Hence, we

---

[8] This is indeed the perspective embraced by Elster (1990: 865). See also note 8.

could say that the knowledge that an action fulfils some notion of common interest suffices to form a set of 'causal' expectation that can be used to sustain a certain social norm on a more stable basis that the sense of resentment would imply. Therefore, public interest is used to explain how motivations can be sustained and expectations can be formed such that they are consistent with the social norm, not as an self-sufficient explanation.

3.6 A Second Reductionist Approach: the Dynamic Process of Convergence Toward a Convention

3.6.1 The Shortcomings of a Static Analysis

The main limitation of the normative expectation theory is, in my view, that it can only be used to account for the *persistence* of a norms *after* this has become established, but it cannot be used to explain how and why this has happened in some more or less recent past. In fact, nearly every outcome of a game could be sustained by means of a 'normative expectations' story, by means of a careful calibration of the weights of the utility function (see section 2.4.1.B, and Chapter 4). Although this is consistent with Robert Sugden's claim that his moral theory follows the tenets of conventionalism, it is needless to say that by doing so many relevant problems are left unanswered.

However, this is also suggestive of a more general problem, relative to the use of a *static* type of analysis; this, in fact, obscures the real individual incentives underlying a convention, since normative expectations, provided that the resentment hypothesis has a sufficient relevance on the individual motivational system, ensure that almost any outcome can be enforced. Therefore, we would need to look at 'what happened before' in order to find out a truly informative account of social norms that abstracted away from normative expectations. What would be needed is then a study of the dynamical pattern that led to the establishment of a convention, i.e. an off-equilibrium analysis. In Chapter 4, then, an alternative model of compliance with norms based on this idea of common interest, where normative expectations are only ancillary to the commitment elicited by this idea, will be developed.

In this Chapter I want to offer a review of the existing theories on social norms that already rely on a dynamical analysis. This is made possible by the approach of evolutionary game theory, which permits to analyse at the same time the equilibrium property of norms and the process of their formation, emergence and persistence.

Let me come back to the concept of ESS that was introduced in section 3.2. I already mentioned how this concept conveyed an evolutionary flavour in that the aspect of 'invasions' from different social norms and 'mutations' of agents' behaviour was taken into account. Now this is explicitly stated and made the central aspect of the account. In fact, the concept of ESS has been carried over from the field of biology to the literature of Evolutionary Game Theory to be applied to a situation in which there exists a large population of agents who are drawn at random at each instant of time and pair-wise matched to play a certain game (Maynard Smith and Price (1973); Maynard Smith (1982)). The agents are 'hard-wired' to play a given strategy, so that the population can be described in accordance to the percentage of players adopting each strategy. At the end of each repetition of the game, however, the average payoff gained by the whole population becomes known to each member, providing the agents with the opportunity to switch their strategy to a more profitable one. However, such an adjustment toward optimality is slow, as agents are considered to be boundedly rational.

A first requirement for a notion of equilibrium in this context would be that agents did not need to change their strategies at any instant of time. This would indeed be suggestive of stability. However, this condition would be trivially satisfied in every situation where, by chance, all the agents played the same strategy and did not have any possibility "to imitate" any other.

3.6.2 The Emergence of a Mutually Beneficial Convention

To give an account of how the Evolutionary approach to norms formation works, let us come back to Sugden's works and let us consider the kernel of the notion of

norm, that of mutually beneficial conventions[9]. Let us consider a particular type of this category, the crossroads game, a symmetrical game that represents a situation of *partial* coincidence of interests. Let us suppose we start from a point of absence of co-ordination among the individuals of the population. The pure strategy Nash equilibria are mutually beneficial, thus satisfying the requirements of mutually beneficial conventions:

	Slow Down	Maintain Speed
Slow Down	0,0	2,3
Maintain Speed	3,2	-10,-10

Figure 3.4

If the game is conceived by the players as *symmetrical* then there is no distinction between being assigned to the Row-player role or the Column player's one, thus imposing that the agents play the same strategy in both situations. In this setting the only Nash equilibrium is the mixed strategy in which "slow down" is played with probability 0,8 and "Maintain Speed" with probability 0.2. Such an equilibrium can be called the "status quo" of the interaction. The result is rather inconvenient: since there exists no convention in assigning the priority at the cross-roads, in the majority of cases people both slow down, and in a minority they both maintain speed, which is the worst outcome for all. Only in 32% of the cases a player gives way to the other. The expected payoff is then 0.4.

Conversely, if the recognition of some asymmetries in the labelling of the players makes the game an asymmetrical one, it is possible to reach equilibria in which agents play different strategies. The Nash Equilibria (Maintain Speed, Slow Down) and (Slow Down, Maintain Speed) are shown to be ESS, while the previous equilibrium in mixed

[9] To be sure, that presented here is only one of the possible approaches to the subject in a growing field of literature. Another, more refined, account is provided by Robert Sugden himself (1998b), who adds to this context a situation of uncertainty over the type of the opposing player. For a comprehensive approach, see Young (1993 and 1998), who puts forward a refinement of the concept of ESS, which must also be robust to random shocks. On questions regarding the stability of equilibria in an evolutionary context, see Skyrms (1997), who argues that in games with more than

strategy fails now to be stable. The basic idea is the following: let us begin from the situation in which each player slows down with probability 0.8; then, suppose that a percentage ε of the agents perceive an asymmetry of whatever kind in the game, say they start thinking that players coming from the right give way to other players. Therefore, they believe that it is convenient for them to maintain speed when coming from the left and to slow down otherwise. If we suppose that the probability of coming from either left or right is the same, such a belief turns out to give them a higher payoff. In fact, when two people of this group of "smart" agents meet each other, they gain a payoff higher than average, which is large enough to compensate for the loss when they meet 'dumb' players. Moreover, when dumb players start to recognise that people maintain speed with higher probability; they are compelled to slow down more frequently. After some adjustments, smart players successfully apply the convention within their group, while dumb players gain the same payoff as before. But the situation is likely to evolve. As soon as larger shares of dumb players recognise that the group of smart players is more successful, they will be willing to shift to the convention. Since the equilibria are mutually beneficial, the group of smart players do not have any reason to prevent them from doing so.

Therefore, the adoption of the convention is likely to propagate to the whole population. Hence, without the presence of any authority, the adoption of a convention increases the welfare of the agents with respect to the status quo: the average payoff is now 2.5. Nevertheless, we cannot predict which of the two alternative Nash Equilibria will be selected: this depends on which direction the initial fraction of mutants will take. In fact, for some agents the final equilibrium could even be worse, for some respect, to its alternative: this is the element of arbitrariness implicit in every convention.

3.6.3 Emergence of Individually Beneficial Conventions

Let us now consider the Hawk-Dove game already introduced in section 3.3. Following the classification offered by Weibull for symmetric games (Weibull (1995:

---

three strategies available to players, the usual mechanism of the replicatore dynamics fails to converge to steady states but generates chaotic trajectories and strange attractors.

40)), this game is formally equivalent to a co-ordination game, thus implying the same type of equilibrium and the same dynamics of convergence toward them. Similarly to the previous game, the equilibrium of war of everyone against everyone in the state of nature is depicted by the equilibrium in mixed strategy when the game is played in the asymmetrical form, not instead by the outcome obtained when everyone is aggressive (H,H), which, would probably give the best representation of the state of war in the state of nature.

Even in this class of games the recognition of an asymmetry helps the development of a regularity of behaviour that spreads across the society and leads to the adoption of the two pure strategy Nash equilibria, which are Pareto superior to the equilibrium in mixed strategy, as occurs in co-ordination games. Sugden seems confident that an asymmetry will be universally recognised: this lies in the *possession* of the thing that people are fighting for, supporting this claim by means of a great deal of empirical evidence.

An analogous process would hold for conventions of reciprocity in the repeated PD, although in this case there seems to be no clear clue as to which rule of behaviour to adhere, since there may be many tit-for-tat like strategies with slightly different forms of punishments, but all would be observationally equivalent in equilibrium. This case then arises some complications from the formal point of view.

3.6.4 Some Critical Remarks

3.6.4.A The Role of Asymmetries

The necessary element that leads to a stable convention is the recognition of some asymmetries in the game. However, it is necessary that players share the same recognition of the asymmetry in order to qualify the game as asymmetrical. Once this happens, then the process evolve spontaneously to a stable convention.

I think that this requirement is rather problematic. Sugden seems confident that such a process of identification of a common asymmetry in the game will eventually emerge: *Sooner or later, (...) some slight asymmetry of behaviour will occur by chance; some players will think that something more than chance is involved, and expect the asymmetry to continue. Even though this expectation has no foundation, it is self-fulfilling* (Sugden (1986: 43)).

However, there are many relevant asymmetries in each game that are likely to attract the attention of each agent. Sugden's answer is based on the reference to Schelling's theory of focal points: some asymmetries are more likely to emerge because of their salience or their prominence. Sugden points out some of the features an asymmetry must have in order to generate a convention:

a) it should be embedded in the structure of the games itself; for instance a distinction between major roads and minor roads in the cross-roads game could offer some appeal, for example because drivers driving on major roads could sometimes fail to recognise a cross-roads with a minor road, thus maintaining speed with higher probability than the previous equilibrium required;

b) if the game is not strictly symmetrical even from the formal point of view, then it is possible that the structure of the payoffs itself could generate a difference in the behaviour of the agents: for example, if the outcome of (maintain speed, maintain speed) gives a slightly better outcome to Player 1 than to Player 2, then the agents who enter the games as Player 1 will have a small incentive to maintain speed with higher probability. When this feature will be recognised by agents who enter the game as Player 2, then a convention in which Player 1 maintains speed and Player 2 slows down is likely to arise;

c) the generality of an asymmetry is very important as well: if an asymmetry is capable to help to indicate a focal point not only in the actual situation in which we are involved, but even in other situations that are analogous for some sort to the current one but differ from it for some other respects, then it is more likely to spread among the population.

3.6.4.B Single and Multiple Focal Points

My understanding of Sugden's argument is as follows: the dynamics of the process leading to the universal adoption of a convention rests on the possibility that agents recognise some sort of asymmetry capable of labelling in some way their participation to the game. This is clearly possible when an asymmetry stands out as an unique focal point, since this will act as a basis for common knowledge helping members of the population to generate a system of concordant mutual expectations. Nevertheless, it is

not possible to rule out the possibility that the "set" of possible asymmetries in the game is multiple. In this case, Sugden relies on a sort of process of trial and error for which the population eventually succeeds in "converging" to recognize a single relevant asymmetry.

Such a process is governed by the argument that a small group of mutant agents experiment new rules of behaviour every now and then. The crucial reason for a population to abandon the status quo and evolve toward a convention is that *within* the group of agents who identifies the asymmetry, it is *convenient* to abide by the rule related to that asymmetry. Since the status quo is given by a mixed strategy equilibrium, then the mutant agents obtain the same payoff as before when matched with agents not part of the group, but a better payoff when meeting some components of the group. Hence, the "innovative" rule of behaviour turns out to be more convenient, even slightly, for the whole population.

However, when the context of the game does not present a prominent asymmetry, the situation is much more problematic. Evolutionary Game Theory actually *assumes* that the group of "mutant" agents adopt the same rule of behaviour. In fact, the assumption that mutations occur once at time is rather simplifying: *evolutionary stability is a robustness test against a single mutation at a time. In other words, it is as if mutations are rare in the sense that the population has time to adjust back to status quo before the next mutation occur* (Weibull, p. 34). This hypothesis could seem justifiable in a biological context, but it appears rather restrictive in a social context, where there are plenty of asymmetries capable of attracting the attention of the agents. In fact, the agents of the mutant group must experiment a problem of co-ordination analogous, if not worse, to the original one, because of the increased number of available alternatives. Thus the original problem seems simply shifted backwards, to the problem of choosing one of the multiple asymmetries that are likely to generate concordant expectations[10]. Of course, this may happen by chance; but the period of time needed to obtain such a homogeneous mutation may well be infinitely long, since its probability is rather small.

[10] On the cognitive problems agents incur in attempting to grasp elements "external" to the interaction, see Sacconi (1986: 171).

The point is that it is necessary an absolute homogeneity in the mutant behaviour if we want the process of evolution to converge to a stable, mutual beneficial outcome. Even in the extreme case of only two, equally "attractive", asymmetries, it is easy to show that no stable convention can emerge: the group of mutant agents, now divided into two subgroups, each adhering to one of the two asymmetries, is no longer better off after the change in behaviour. They will experience a worse payoff playing with others mutant agents, since there is now the possibility to be matched with a mutant agent of the different subgroup. Thus the spread of the new rule is hindered from the beginning.

This problem is far more relevant if we introduce some changes in the formal structure of Evolutionary Game Theory to "adapt" it to the context of social interactions. Rather than thinking that a mutation in the general rule of behaviour by a restricted part of the population happens by chance, we could, more realistically, assume that this change is *voluntarily* brought about by a minority of "smart" agents, capable of reaching a certain degree of understanding of the complex of the interaction. In this way, the periodic mutations would not be a random process, but the result of a conscious evaluation by "enlightened" agents. Now, in these circumstances the smart agents will recognise that every change in their behaviour is detrimental for them, unless all of them happen to identify the same asymmetry in the game, an event that has a too small probability to be considered. The result would be that no more mutations in behaviour would be brought about, thus obstructing from the beginning the evolutionary process that should lead to the general adoption of a convention.

In this context, it would seem sensible to argue that the smartest agents simply *agree* on the choice of a certain asymmetry as relevant in order to solve their co-ordination problem. This identification of this asymmetry will become common knowledge between their "club", on the basis of their agreement. As the equilibrium is mutually beneficial, then the group of smart players will be better off if other players join the club. As the process goes on, the club of smart players will extend to the whole population. This process would follow the same dynamics as that depicted by Sugden, with the peculiar difference that the basis for the common knowledge is now

given by an agreement rather than by the prominence of an asymmetry, thus offering a sounder basis as a device to generate concordant and mutual expectations and fostering a quicker convergence toward the general adoption of a convention.

3.6.4.C Possession in Property and Reciprocity Conventions

As a further application of the argument above, I will now give some critical remarks about the possibility of emergence of a rule of assigning property rights. The relevant asymmetry indicated by Sugden was that of the possession of the thing contended. Clearly, the question as to how this allocation of possession has been reached is left unanswered. We could imagine that a precedent situation of conflict for the possession of things arose in the state of nature, but this situation would share the same features of that just analysed in the conflict for the property of things. Therefore, we could think of a *preliminary* hawk-dove game to solve the conflict over possession, and then a second stage of the same game for the *property* of things. But this means that the crucial problem of the war in the state of nature is, again, simply shifted backwards.

Alternatively, we could think that there is a somehow universally shared *rule* to assign possession of things, that does not bring about any situation of conflict, upon which the property games can be sorted out. But what could such a rule consist of? Could it be represented by a concept of *geographical proximity* toward the thing object of the conflict, as many examples of Sugden's analysis could let us think of? But in this case a conflict for the possession of the best and richest area of the territory is likely to arise. Perhaps we could think that a somewhat *impartial* rule, that everyone could accept with no controversy, could be adopted as a device to solve this problem. But this would be antithetical to Sugden's approach: impartiality is in itself a moral concept, thus it should be the outcome of a process of evolution, not its starting premise. The same sort of argument applies to the reciprocity conventions as set out in section 3.2.3 In that case a convention was seen as one of the possible tit-for-tat-like strategies with different periods of punishment. In this case it is even harder to perceive a remarkable asymmetry capable of triggering the process of convergence to a stable outcome.

CHAPTER 4

THE EVOLUTION OF NORMS AND INDIVIDUAL CHOICES

INTRODUCTION

Two main critiques of the concept of normative expectations have been put forward in Chapter 3. The first refers to their role as self-enforcing devices for conventions; in this respect, I have argued that not paying attention to the dynamical process that gets a convention established risks basing the entire analysis on *ad hoc* assumptions. The second critique refers to the dynamical account offered by Sugden, in which the requirement of some relevant asymmetry in a game appears to beg the question of the emergence of conventions, in that it shifts the burden of the explanation just one step back.

In this chapter I would like to focus on the first of these criticisms. The argument will be restated by means of a dynamical model in which social norms and individual preferences are jointly analysed within a unified framework. I hope that by doing so, what I perceive to be the main shortcomings of normative expectations theory, at least in the particular specification that I consider, will appear clearly. Such a model will be contrasted with an alternative model of individual motivation built along the lines of the analysis developed in Chapter 2.

However, I believe that, along with the substantive contents of the theories to be compared, another- and perhaps more fundamental - element of novelty in the present chapter is given by the methodological framework that is developed. As stressed in the introduction, the final goal is to tackle the two-sided relationship between individual behaviour and social norm in a comprehensive framework. In order to do this, we need to adopt two basic tools. The first is the multiple-motivations utility function developed in Chapter 2, which enable us to study how social norms shape individual

motivations. The second is a dynamical analysis that permits to detect how individual action modifies aggregate patterns of behaviour, which, in equilibrium, give rise to social norms. In order to fulfil this latter requirement, I shall draw on the evolutionary approach outlined in section 3.6, whose main assumptions are the bounded rationality of individuals, and the process of change of individual actions on the grounds of their degree of 'fitness' with respect to the environment. In particular, the 'replicator' dynamics implies that individual strategies that have a better fit with the environment spread across the population of agents through a process of imitation.

The evolutionary approach is certainly well suited to study the process of change of aggregate patterns of behaviour that shapes social norms. In fact, the assumption of bounded rationality can be thought of as deriving from the adoption of a short-term perspective, in which, because of time constraints, agents have not yet acquired the information necessary for a fully informed choice and the ability to fully decipher the environment they are acting in. Since a dynamical approach necessarily adopts a very short time-unit, at least for the type of situation we wish to describe, the adoption of a boundedly rationality approach seems indeed sensible. This argument in support of the evolutionary theory seems in fact to gather consensus amongst many economists, even amongst those in favour of the full rationality approach (see e.g. Lucas (1986)). For instance, Mailath (1998) states explicitly that the only epistemological account of the concept of Nash equilibrium is that relying on the evolutionary paradigm and on the process of learning by individuals.

Hence, a general framework for the dynamical analysis of norms and individual behaviour will be developed in the present Chapter, which draws on evolutionary game theory instruments on the one hand and the multiple motivations model of individual utility on the other. A generalisation of the notion of Psychological Nash equilibrium to the present setting will be put forward in section 4.1. This is then applied in section 4.2 to the study of the Prisoner's Dilemma for the case in which agents have normative expectations. In section 4.3 two 'dynamical' concepts are developed, based on two different ways of modelling off-equilibrium beliefs. These are then applied to the same setting as before, highlighting what I consider to be some shortcomings in the theory of normative expectations. Section 4.4 takes on an

alternative model of other-regarding motivations, which are shaped as conditional willingness to comply with a shared view of morality. It is shown how this model is not subject to the shortcomings of the previous one.

4.1 THE GENERALIZATION OF THE NASH PSYCHOLOGICAL EQUILIBRIUM TO CONTINUUMS OF POPULATIONS

In the present Chapter I shall take Sugden's model of individual choice put forward in section 2.4.1.B as illustrative of the normative expectations theory, although some other specifications are possible, such as the rather more complex one that Sugden himself has used in other works (1998b). In my view, the former, besides being amenable to formal analysis unlike the second, appears capable of capturing the main insight of normative expectations.

Recall from section 2.4.4.B and 2.3.3 that this model fits the general version of a utility function divided into a self-interested and an other-regarding motivation represented as follows:

$$V_i(\sigma;b)=U_i(\sigma,b)+\lambda_i f[T(\sigma,b)] \tag{2.5}$$

The other-regarding motivation is given by the following function, which incorporates the resentment hypothesis:

$$f(T(\sigma);b)=\begin{cases}0 & \text{if } m(\sigma_i;p_i,p_j)\geq 0\\ m(\sigma_i;p_i,p_j) & \text{if } m(\sigma_i;p_i,p_j)<0\end{cases} \tag{2.7}$$

The *impact* function *m(.)* expresses the difference between the *actual* payoff accrued to player *j* by *i*'s action and *j*'s *expected* a-priori payoff.

$$m(\sigma_i;p_i,p_j)=U_j(\sigma_i;p_j)-U_j(p_i;p_j) \tag{2.6}$$

Also, recall that in the presentation of the model I assumed that there were two continuums of agents, whose members were drawn at random to enter the game in the two different positions. p_i and p_j represent the vectors of *average play* for the *i*-players and *j*-players populations, respectively; that is, for a given $l=i,j$:

$$p_l(s_l)=\int_{\sigma_l\in\Sigma_l} P_{\sigma_l}(s_l)\vartheta(\sigma_l)d\sigma_l \tag{4.1}$$

where $P_{\sigma_l}(s_l)$ is the probability of playing the pure strategy $s_l \in S_l$ according to the mixed strategy σ_l, and $\vartheta(\sigma_l)$ is the *density* of players using the mixed strategy σ_l, which satisfies the condition $\int_{\sigma_l \in \Sigma_l} \vartheta(\sigma_l) d\sigma_l = 1$; that is, the integral over all the strategies densities exhausts the Leabesgue-measure of the whole population, which has been conventionally set equal to 1. Hence, $U_j(p_i, p_j)$ is the payoff that an *i*-player gauges a *j*-player is expecting, given the common knowledge on average plays. $U_j(\sigma_i; p_j)$ is instead the *actual* expected payoff accrued to a *j*-player by the *actual* play by agent *i*, i.e. σ_i. The difference between the two is a measure of the resentment elicited in the *i*-player by lowering the *j*-player's payoff from what expected.

Since the average plays p_i and p_j are common knowledge amongst players, I assume that individual beliefs are consistent with them; that is, $b_i = \beta(p_i, p_j)$. This also permits a simplification of the notation: the comprehensive utility function will generally be indicated as a function of average plays, rather than beliefs:

$$V_i(\sigma; b) = V_i(\sigma; \beta(p_i, p_j)) \equiv V_i(\sigma; p_i, p_j) \qquad (4.2)$$

Hence, the first argument of $V(.;.)$ refers to actions, whereas the second refers to the average plays on which beliefs are formed.

The first issue we have to address concerns the set of Nash Psychological equilibria of the game. Recall from section 2.3.3 that this consists of two conditions: consistency of expectations with equilibrium play, which in the present setting is ensured by the just mentioned condition (4.2) and individual optimality, which amounts to agents not having an incentive to deviate from the prescribed equilibrium behaviour. It seems natural to add a further condition, which requires coherence between individual and aggregate behaviour; that is, optimal individual behaviour should coincide with the average play within the population. If this condition did not hold, in fact, every agent would find it optimal to perform a different behaviour than the average, which would cause an obvious inconsistency between average and individual behaviour. In other words, an average behaviour that did not reflect optimal behaviour at the individual level would be likely to be swept out by a process of

adjustment of players toward optimality, which is assumed to be possible, although at a relatively *slow* rate, within an evolutionary setting.

Taking account of the notation introduced in this section, a Nash Psychological equilibrium can be restated as follows: it will be given by a pair of average plays $(\hat{p}_i, \hat{p}_j)$ such that:

i) $\hat{b} = \beta(\hat{p}_i, \hat{p}_j)$

ii) for each $l \in I$, $\hat{\sigma}_l$ that satisfies $V_l(\sigma_l; \hat{b}) \leq V_l(\hat{\sigma}_l; \hat{b})$ for every $\sigma_l \in \Sigma_l$, (4.3)

is such that $\hat{\sigma}_l = \hat{p}_l$

Notice that the overall set of players is given by the union of the two populations; that is, $I = I_i \cup I_j$. Hence, subscript l refers to a generic player in either population. Condition (i) is based on the account of consistent beliefs introduced in section 2.3.3 and ensures the coherence of individual beliefs with average behaviour in the populations. Condition (ii) ensures that individual behaviour is optimal and that average behaviour coincides with individually optimal behaviour.

4.2 Static Psychological Nash Equilibria in a Prisoner's Dilemma

4.2.1 Substitute Strategies

In applying this model, I shall focus on the following general version of the Prisoner's Dilemma, where the limitations that $\beta > \gamma > \alpha > \delta$ ensures the fulfilment of the usual properties of the interaction:

	Co-operation	Defection
Co-operation	γ,γ	δ,β
Defection	β, δ	α,α

Figure 4.1

For the purpose of the analysis, it will become key whether the quantity $(\beta-\gamma)$ exceeds $(\alpha-\delta)$. Let us first assume that

$$\eta \equiv (\beta-\gamma)-(\alpha-\delta)>0 \tag{4.4}$$

Making use of a definition put forward in the literature (Fershtman and Weiss (1998)), under condition (4.4) individual strategies can be called *substitutes*, as the 'disincentive' to Co-operate is larger when the other party is Co-operating than when she is Defecting. In the final part of the section I shall illustrate the contrasting case of *complementary* individual strategies. Let us denote the percentage of co-operators in either population with p_i. Now consider the situation of a generic player i, who knows (and knows that it is common knowledge) that the percentage of plays in either population is given by the pair (p_i, p_j). First, she has to compute the expected payoff for a generic j-player, on the grounds of the first and second order beliefs consistent with the pair (p_i, p_j). This will be given by the following expression:

$$E_i[U_j(p_i, p_j)]=(\gamma+\alpha-\beta-\delta)p_i p_j+(\beta-\alpha)p_i-(\alpha-\delta)p_j+\alpha \tag{4.5}$$

Consequently, the impact function for player i by playing σ_i is:

$$m_i(\sigma_i; p_i, p_j)=[(\gamma-\delta)p_j+(\beta-\alpha)(1-p_j)](\sigma_i-p_i) \tag{4.6}$$

Notice that the sign of m_i only depends on the sign of the expression $(\sigma_i - p_i)$. Other-regarding utility can thus be rewritten as:

$$f(\sigma_i; p_i, p_j)=\begin{cases}0 & \text{if } \sigma_i \geq p_i \\ [(\gamma-\delta)p_j+(\beta-\alpha)(1-p_j)](\sigma_i-p_i) & \text{if } \sigma_i < p_i\end{cases} \tag{4.7}$$

This expression is consistent with the resentment hypothesis as modelled by Sugden (1998a), in that implementing a co-operative action with higher probability than the average does not provide a higher payoff; the opposite is true when a less co-operative action is performed.

The overall extended utility for agent i is then given by:

$$V_i(\sigma_i; p_i, p_j)=(\gamma+\alpha-\beta-\delta)\sigma_i p_j+(\beta-\alpha)p_j-(\alpha-\delta)\sigma_i+\alpha+\lambda_i f(\sigma_i; p_i, p_j) \tag{4.8}$$

We now have to work out what is the optimal action for agent i. This can be done by differentiating expression (4.8) with respect to σ_i, which leads to:

$$\frac{\partial V_i(\sigma_i; p_i, p_j)}{\partial \sigma_i} = (\gamma + \alpha - \beta - \delta)p_j - (\alpha - \delta) + \lambda_i[(\gamma - \delta)p_j + (\beta - \alpha)(1 - p_j)]Ind(\sigma_i < p_i) \tag{4.9}$$

where

$$Ind(\sigma_i < p_i) = \begin{cases} 1 & \text{if } \sigma_i < p_i \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

The first insight is that it is never optimal for agent *i* to perform a 'super-erogatory' action: if $(\sigma_i - p_i) > 0$, then the latter term of the differential is nil, whereas the first two are both negative. This of course depends on the way normative expectations are shaped, as in particular they do not reward with a positive extra utility actions that accrue *more* utility than expected to the opponent. In other words, only social disapproval affects negatively one's overall utility, whereas social approval leaves the agent indifferent or is not relevant as far as motivations are concerned.

Let us then focus on strategies such that $(\sigma_i - p_i) \leq 0$. By solving equation (4.9), the following inequality obtains:

$$\frac{\partial V_i(\sigma_i; p_i, p_j)}{\partial \sigma_i} \geq 0 \Leftrightarrow p_j \leq \bar{p}_j \tag{4.11}$$

where

$$\bar{p}_j = \frac{\lambda_i(\beta - \alpha) - (\alpha - \delta)}{(1 + \lambda_i)\eta} \tag{4.12}$$

In order for inequality (4.11) to be strategically meaningful, namely, to make $\bar{p}_j$ lie between zero and one, we must impose some restrictions on the parameters; in particular, we want λ_i to lie at 'intermediate' levels. In fact, if λ_i is too high, then the weight attached to other-regarding utility is too high and the individual will always Co-operate no matter what the other party is doing. If λ_i is too low, the agent will never Co-operate and will act as a purely self-interested individual. Therefore, λ_i must satisfy the following constraint in order not to have a trivial situation:

$$\lambda_{\min} = \frac{(\alpha - \delta)}{(\beta - \alpha)} < \lambda_i < \frac{(\beta - \gamma)}{(\gamma - \delta)} = \lambda_{\max} \tag{4.13}$$

In fact, if λ_i did not lie at 'intermediate' levels, then it would make either unconditioned Co-operation (when λ_i is relatively high) or unconditioned Defection (when λ_i is relatively low) the dominant strategies for the agent. Throughout the Chapter, instead, I shall focus on those cases that are strategically more interesting and that do not prescribe an unconditioned behaviour to an agent. More precisely, conditions (4.13) concern the *inclination to resentment* of an individual when failing to live up to others' expectations; overall, they state that resentment will be the prevailing motivation only in the context that is *less* costly in terms of self-interest. Since in the present context of substitute individual strategies, Co-operation is *more* costly when the other party is *Co-operating* rather than when she is Defecting, resentment will *permit* Defection when the counterpart is Co-operating, and will *impede* Defection when the other party is Defecting[1].

This explanation should also make it clear the rationale of condition (4.13); under the substitute strategies assumption, the probability with which the opponent, on average, Co-operates must not be too high in order to spur the Co-operation of agent *i*; in fact, were it too high the individual would start to Defect, as in that case the self-interested motivation overcomes resentment considerations. Conversely, if the opponent Co-operates with a sufficiently low probability, the inclination to resentment will trigger a Co-operative behaviour. Obviously, given the symmetry of the game, an analogous condition holds for *j*-players.

---

[1] In fact, the first inequality can be rearranged to yield: $\lambda(\beta-\alpha)>(\alpha-\delta)$
The first term is the loss, due to resentment, of other-regarding utility, whereas the second is the benefit in terms of self-regarding utility stemming from a drop in the probability of co-operation, provided that the other party is defecting. Therefore, this condition ensures that the resentment cost outstrips the self-interested benefit under defection from the other party. Analogous considerations hold for the second inequality, which can be so re-expressed: $\lambda_i(\gamma-\delta)<(\beta-\gamma)$

Here, the first term represents the resentment for failing to co-operate and the second the self-interested gain, provided that the other party is co-operating.

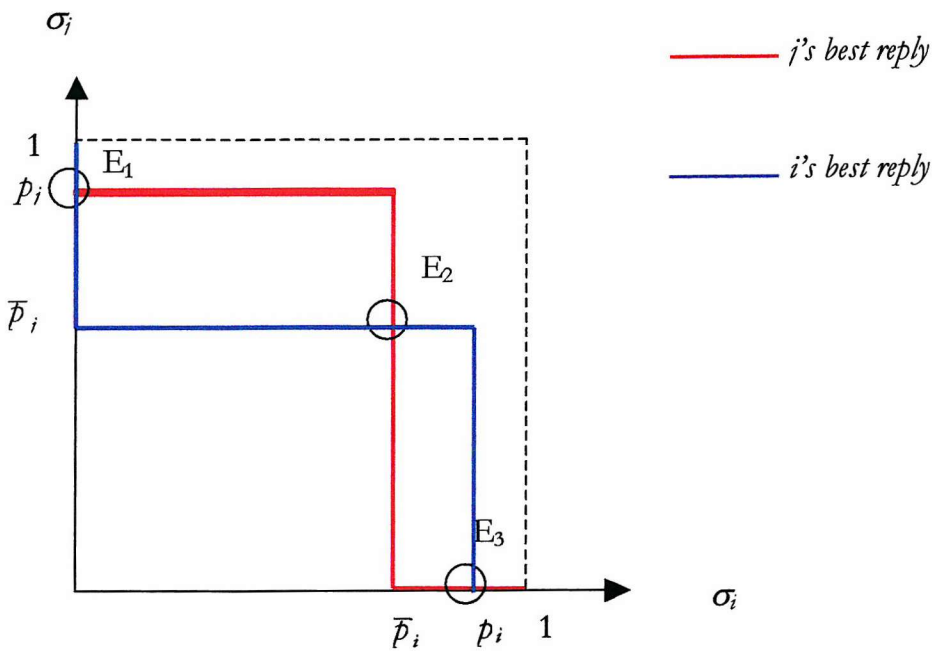

Figure 4.2

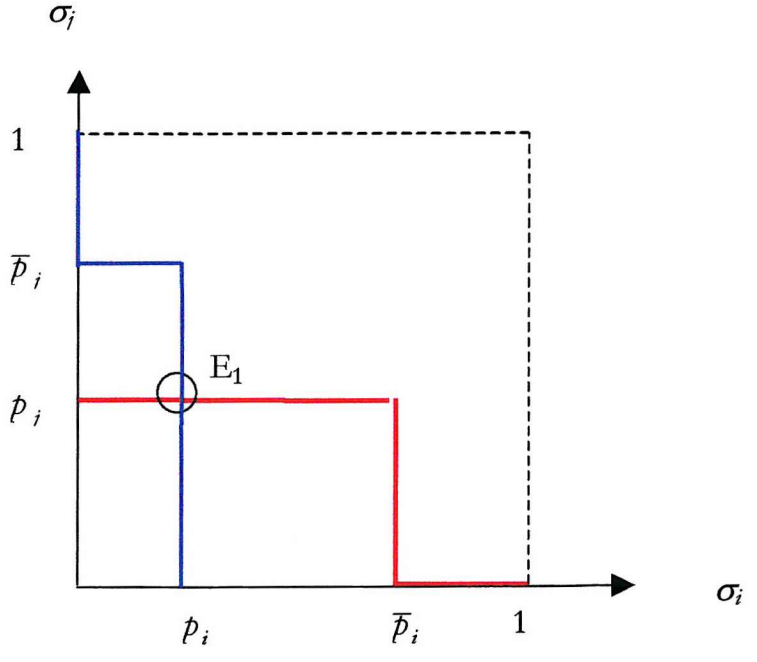

Figure 4.3

Graphical analysis shows that a large number of equilibria is possible. In Figure 4.1, I have depicted the best reply functions for two generic players belonging to the *i*-population and *j*-population. Notice that the two threshold levels are not necessarily the same, as they could differ for a different value of λ_i, i.e. the two populations may be different because of the weight attributed to other-regarding utility. Moreover, the shape of the function is such that it is never optimal to Co-operate with higher probability than the average of the population; that is, the best reply function for player *i* is constrained to lie below p_i. A preliminary condition to find an equilibrium is that, as usual, the two best reply functions intersect. However, this is not enough, as condition (4ii) also states that individual optimal play must coincide with average play in a population. Therefore, none of the three candidates for equilibrium circled in Figure 4.2 can be considered equilibria of the game. Instead, outcomes in which the population average play is below the threshold level determine an equilibrium for the game. Such are the configurations belonging to the set: $E_1 = \{(\hat{p}_i; \hat{p}_j): \hat{p}_i \leq \overline{p}; \hat{p}_j \leq \overline{p}\}$. Figure 4.3 shows one of such equilibria:

Point E_1 in Figure 4.3 is a mixed strategy equilibrium, where the probability of Co-operation is bounded from above by the two threshold levels. This makes the corresponding outcome overall *inefficient,* in the usual sense in which mutual Defection is inefficient in a PD. Moreover, since no agent is required to produce a super-erogatory action when the other agent is not, a situation that will be analysed below, we may qualify this set of outcomes as *reciprocal.* In fact, the probability of Co-operation is low because expectations on each other population's Co-operation is low, which fails to trigger the resentment mechanism. Hence, such a set can be called an *inefficient reciprocal* type of equilibria. Notice that it also includes as a particular case the standard Nash equilibrium of the game $(\hat{p}_i = 0; \hat{p}_j = 0)$.

Figure 4.4 shows a type of equilibrium where all the agents of a population act submissively – namely, they co-operate with high probability - whereas all of the others act exploitatively. As the picture shows, all the outcomes such that $E_2 = \{(\hat{p}_i; \hat{p}_j): \hat{p}_i = 0; \hat{p}_j \geq \overline{p}\}$ are equilibria (of course, symmetrical outcomes are equilibria as well). To mark the contrast of this set of equilibria with the other, I shall

call this type *anti-reciprocal*, or *exploitative*, in that one group of individuals is prompted to Co-operate by the very fact of others' Defection: on the one hand, resentment-inclined individuals will feel obliged to live up to *i*-players expectations, demanding as these may be. On the other hand, the very low level of expectations set on *i*-players in relation to their Co-operation, justified by their population's general opportunistic behaviour, suffices to avoid the resentment of their opponents. The seemingly paradoxical character of this equilibrium lies in that it is sustained on the grounds of expectations that, recalling previous definitions, are *empirical* but not *causal*; that is, general conformity to the Co-operative norm by *j*-players is not triggered by considerations in terms of self-interest, but from the mere past conformity of individuals in that population. We shall expand on this point later on.

4.2.2 Complementary Strategies

So far, we have obtained an *inefficient reciprocal* set of equilibria, characterized by the set E_1, and an *anti-reciprocal* one. Under the hypothesis of *complementary individual strategies*, though, a set of *(almost)-efficient* equilibria is possible. This holds under the following condition:

$$\eta<0 \tag{4.14}$$

As a result, the previous optimality inequality (4.11) is now reversed:

$$\frac{\partial V_i(\sigma_i; p_i, p_j)}{\partial \sigma_i} \geq 0 \Leftrightarrow p_j \geq \overline{p}_j \tag{4.15}$$

where the threshold value is the same as in expression (4.12). Now, the conditions that ensure the feasibility and the non-triviality of (4.15) are as follows:

$$\frac{(\beta-\gamma)}{(\gamma-\delta)} < \lambda_i < \frac{(\alpha-\delta)}{(\beta-\alpha)} \tag{4.16}$$

The interpretation is the same as that outlined above; only, individual strategies being complements determine a reversal of the terms of those inequalities. This third type of equilibria is illustrated in Figure 4.5. This set can be given a general representation as follows: $E_3 = \{(\hat{p}_i; \hat{p}_j): \hat{p}_i \geq \overline{p}_i; \hat{p}_j \geq \overline{p}_j\}$. The economic intuition is analogous, but 'opposite in sign' with respect to that given for E_1 and E_2.

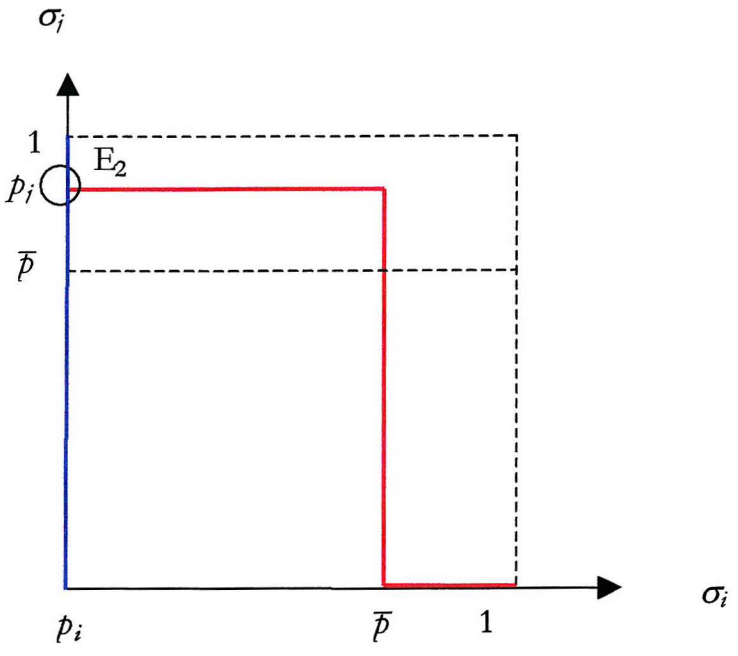

Figure 4.4

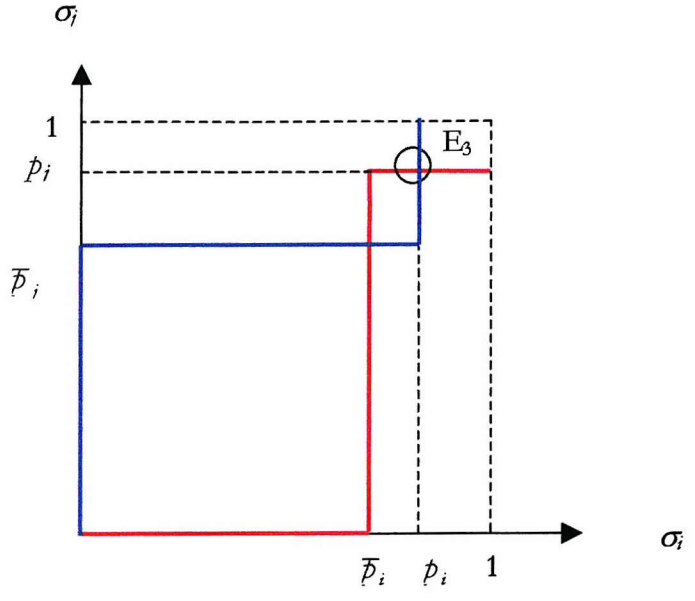

Figure 4.5

The inclination to resentment is now triggered when the other party Co-operates, given the smaller opportunity cost borne by the individual in this situation. Therefore, each individual has sufficient incentive to Co-operate when the other is co-operating, thus bringing about this *reciprocal* equilibrium. Since the probability of Co-operation is now bounded from below, it seems natural to call this an *efficient,* or *almost-efficient,* equilibrium. In this setting, no equilibrium can be sustained such that agents co-operate with probability less than $\overline{p}$, the only exception being the standard Nash equilibrium where both populations always Defect.

4.3 The Dynamics of Normative Expectation

4.3.1 How to Interpret the Replicator Dynamics?

The presence of both a vast number of equilibria, and of different *types* of equilibria in the case of substitute individual strategies makes the question of their stability even more interesting than usual. The issue I wish to address is whether such equilibria are stable from the dynamical point of view. This requires defining two conceptual tools. The first is a plausible model of dynamical evolution of the agents' behaviour. The second is a concept of equilibrium, and of stability, in the dynamical setting.

As for the first, I shall adopt the replicator dynamics as a rule of motion of agents' behaviour. Given the extensive studies carried out on the properties of replicator dynamics, all the pros and cons of its application are well-known (see Weibull (1995); Young (1998)) Though, I should at least spend some words on its suitability for the case under study. In fact, the application of replicator dynamics to social interactions is usually justified on the grounds of the 'parable' of the *imitation of most successful agents.* That is, individuals adopt the strategies used by other agents once they realise that these bring about better results than the strategies they are currently using. The adjustment to the currently more profitable strategies is not immediate, as information does not spread instantaneously through the system, and because agents are not always able to process that information in the most profitable way. This is why replicator dynamics can be considered an aggregate model of evolution responding to the behaviour of boundedly rational agents. Behind this general justification for the

employment of replicator dynamics, there lie some more specific underpinnings. First, better micro-founded accounts of this dynamic process can be offered. Second, other processes of evolution can be shown to lead, under some conditions, to the same results in the long run.

The more controversial issue concerning *all* evolutionary criteria, not merely replicator dynamics, regards *what* is to be understood by 'success'. In many contexts this has a clear connotation, e.g. profit, for firms involved in a competitive market. In others, however, two inter-related aspects contribute to make this issue rather problematic. First, there is the question of identifying the *individual* notion of success. This directly recalls the analysis on reasons to action that I outlined in the previous sections, since the more natural definition of success, in a social context, lies in a subjective concept like *preference satisfaction.* Even if this issue is disentangled by opting for a subjective account of value, as we did in section 2, however, the relevant problem becomes that of how can such a subjective concept of success be taken as a reference point in modelling imitation between agents. In other words, it seems problematic to argue that an agent will embrace someone else's theory of value because its fulfilment brings about greater 'satisfaction'. This is nothing but the issue of inter-comparability, or inter-subjectivity, which was raised in Chapter 1.

To make matters even more complicated in the present context, there is the fact that not only does each individual have a subjective account of values, but also one of its component is given by resentment on others' expectations, which seems an even less tangible element than subjective self-interest. A strategy that some scholars adopt is to apply replicator dynamics to the self-regarding rather than to the other-regarding component; that is, individuals' behaviour carrying greater 'material' or 'economic' success diffuse more rapidly across the population, unlike the fulfilment of their other-regarding motivations (Fershtman and Weiss (1998)). This account seems consistent with the biological idea of 'success' as 'fitness with respect to the environment', which in a social context would find its more direct counterpart in some economic standards. However, those scholars' argument seems in some way to beg the question as they assume the possibility of recognizing one another's disposition to Co-operate, thus

indirectly making the socially-rewarded behaviour the most successful one in 'fitness' terms.

Notwithstanding all these *caveats*, in what follows I will still adopt the replicator dynamics as the basic evolutionary mechanism. Besides representing the most tractable model to describe the motion of aggregate behaviour, in fact, some of its shortcomings set out above can perhaps be better defended. I will not deal with the issue of imitation with respect to a subjective account of value, as this is a general problem that applies to dynamical criteria based on any idea of learning. Rather, the other issue concerning the possibility of imitating behaviour that engender less resentment than others seem, perhaps surprisingly, less problematic in that in this case an 'objective' criterion to allow comparison between agents, that is, resentment, does exist. This is at least the position taken by Sugden when he advocates a 'naturalistic' account of moral theorizing and argues that resentment is a basic motivational force of human beings.

The same idea can be restated in somewhat different terms by arguing that 'social status' is a better (more objectively) definable characteristic than individual success, as it is generally conceived as being related to an inter-subjective source of value. Accordingly, general esteem is accorded to behaviours considered socially virtuous, and these spread through the population by means of emulation.

Another, apparently more technical, issue concerns the use of mixed strategies at the individual level, as I assumed in the previous analysis, despite most works have been carried out under the assumption of agents performing only pure strategy. As will be immediately clear, this latter choice makes the analysis easier under many respects. However, as highlighted by Fudenberg and Levine (1998), this is not a neutral choice as dynamics based on pure strategy seem to have a 'stabilising' effect in some cases with respect to a mixed strategy dynamical mechanism. In what follows I will still put forward a basic definition allowing for agents using mixed strategies, thus making the analysis comparable to that carried out in the static context. I apply a qualitative investigation of the properties of the equilibria in the rest of the section.

4.3.2 Deviations with Steady State-Consistent Beliefs: The GPS Replicator Steady State

The original notion of Nash Psychological equilibrium presented by GPS (1989) only holds in a static context. Besides their basic definition (reproduced in section 2.3.3), they also put forward some refinements of this concept with the purpose of carrying over notions such as that of *(trembling-hand) perfect* equilibria to the new setting. The key characteristic of this type of refinement is that equilibrium strategies are slightly perturbed, thus allowing for any other strategies to be played with an arbitrary small probability (Myerson (1991)). A static equilibrium is then said to be *trembling-hand perfect* if it is still an equilibrium for all the 'perturbed' games as the perturbation becomes increasingly small. One can then interpret such a concept as making the equilibrium *robust* to small changes in the related strategy, where such changes, in some sense, 'converge' to it; hence, some unsophisticated conception of dynamical stability can be said to be embedded in this concept[2]. Therefore, it is possible to start from here in order to develop a notion of stability in a dynamical setting.

In the Nash Psychological equilibrium, the main characteristic of these refinements is that 'off-equilibrium' beliefs are required to be coherent with the equilibrium strategy. That is, even on off-equilibrium paths it is common knowledge that average play is consistent with that played under the GPS Nash equilibrium. In fact, once this notion is carried over to the present dynamical setting, its rationale is that what is being tested is whether the behaviour of players whose *Leabesgue-measure* is negligible with respect to the whole population, will converge or not, once a set of players whose Leabesgue measure is equal to 1 – namely, to the measure of the entire set - are actually playing the static *equilibrium* strategy. Only in this case would it be plausible to assume common knowledge of the would-be equilibrium strategies when analysing the situation *off* the equilibrium. In other words, this notion of dynamical equilibrium investigates the robustness of the equilibrium as changes by very 'few' mutants within

---

[2] In reality, what still makes this notion a static one is that the perturbed games are at any rate considered in isolation from each other; that is, even if any equilibria of a 'succession' held separately from each other, it still would not imply that there was a 'tendency' for the play to become 'attracted' by the equilibrium play. One could conclude that in this case there exist an analogous relation to that between evolutionary stable strategies and stable steady states of a replicator dynamics.

the population occur, while the bulk of the population stick to the 'candidate-to-equilibrium' strategy. In the next section, I shall discuss a stronger notion of stability, where deviations by sub-sets of the population that have positive measure are allowed.

Since we are dealing with mixed strategies, the replicator equation needs some amendments with respect to its standard version. Recalling notation introduced in section 4.1.1, its application to each density yields:

$$\frac{\dot{\vartheta}(\sigma_l)}{\vartheta(\sigma_l)} = V_l(\sigma_l; p_l) - \overline{V}_l(p_l) \tag{4.17}$$

where $\overline{V}$ is the average payoff obtained in population l:

$$\overline{V}_l = \int_{\sigma_l \in \Sigma_l} V(\sigma_l)\vartheta(\sigma_l)d\sigma_l \tag{4.18}$$

Notice that equation (4.17) and (4.18) leave unaltered the overall measure of the population over time. If one wanted to calculate the change in the play of a pure strategy, then, one should keep track of the changes in every density:

$$\dot{p}(s_l) = \int_{\sigma_l \in \Sigma_l} P_{\sigma_l}(s_l)\dot{\vartheta}(\sigma_l)d\sigma_l \tag{4.19}$$

A GPS replicator steady state can then be defined as a vector $\hat{p}_l$ such that

(i) $\hat{p}_l$ is a solution to the system $\dot{p}(s_l) = 0$

(ii) In all equations (4.19) – (4.21): $V_l(\sigma_l; b_l) = V_l(\sigma_l; \beta(p_l))$ (4.20)

Condition (4.20 i) is the standard notion required for a steady state. Condition (4.20 ii) holds that beliefs are consistent with $\hat{p}_l$ itself. However, in the two-strategy case with which I shall be dealing with in the following sections, it is easier to look for the solution to the system of differential equations (4.17) instead of that formed by (4.19):

(i') $\hat{p}_l$ is a solution to the system $\frac{\dot{\vartheta}(\sigma_l)}{\vartheta(\sigma_l)} = 0$ *for any* $\sigma_l \in \Sigma_l$ (4.21)

In fact, this is a *more* restrictive condition than the previous one. It requires that in equilibrium there is no tendency for any mixed strategy to change its frequency, as they all fare the same as the average play given by $\hat{p}_l$.

That players have no incentive to change their mixed strategies does not necessarily imply that the associated steady state is stable; indeed, stability requires the

tendency of the system to converge on - or not to move far away from - the steady state position, after some variables have been perturbed. This usually straightforward notion here requires some qualifications as we have two types of 'variables' that are qualitatively different: strategies and beliefs. In other words, we need a condition telling us how beliefs are shaped *off-equilibrium.* The answer that seems in line with GPS original paper is possibly the simplest one: beliefs are consistent with the steady state average play. No argument, other than analytical simplicity, is offered in GPS to underpin this hypothesis. As suggested earlier, this specification is coherent with the idea that *deviations* from the steady state equilibrium are performed by a set of agents whose Lebeasgue-measure is zero.

Rather than considering the mathematical notion of *local* stability of a steady state based on the theory of linear systems of differential equations, I will find it easier, and also more appealing from the intuitive point of view, to deal with the following *analytical* notion, especially in the two-strategy case, to which the following condition refers:

A GPS replicator steady state $\hat{p}_l$ is *Liapunov-stable* if, besides satisfying (4.21) and (4.20 ii), it also fulfils the following condition[3]:

$$\exists\ \omega > 0 \quad \text{s.t.} \quad \forall\ \sigma_l \in |\sigma_l - \hat{p}_l| < \omega \quad \{V(\sigma_l;\hat{p}_l) - \overline{V}(\hat{p}_l)\} \leq 0 \qquad (4.22)$$

Notice that the first term of the last inequality is that determining the growth rate in the frequency of a strategy σ_l. Therefore, this condition implies that strategies *above* $\hat{p}_l$ are characterised by payoffs no more than the average, so that the relative frequency will not increase over time, and *vice versa.* Overall, then, frequencies are such that they will *not diverge* with respect to the steady state frequency $\hat{p}_l$. In particular, the fact that the main inequality of (4.22) can also be satisfied with equality means that it suffices that the system is not *led away* from the steady state, but it cannot guarantee that the system comes closer to it either. This is why I have labelled the previous concept 'Liapunov' stability, as such a concept indeed only requires the system not to 'go away' from the steady state (see Hirsch and Smale (1974)).

---

[3] The generalisation of this condition for the n-strategy case would be as follows:
$\exists\ \omega > 0 \quad \text{s.t.} \quad \forall\ \sigma_l \in \|\sigma_l - \hat{p}_l\| < \omega \quad , \forall\ k = 1..n, \ \{V(\sigma_k^l, \hat{p}_{-k}^l;\hat{p}_l) - \overline{V}(\hat{p}_l)\} \leq 0$

If instead we wanted to add the strongest condition that the system *does* converge toward the steady state, then the main condition of (4.22) should hold with *strict* inequality. In this case, I shall talk of *local asymptotical stability:*

A GPS replicator steady state $\hat{p}_l$ is said to be *locally asymptotically stable* if, besides satisfying (4.21) and (4.20ii), it also fulfils the following condition[4]:

$$\exists\ \omega>0 \ \text{ s.t. } \ \forall\ \sigma_l \in |\sigma_l - \hat{p}_l| < \omega \ \ \{V(\sigma_l;\hat{p}_l)-\overline{V}(\hat{p}_l)\} < 0 \qquad (4.23)$$

Now, strategies *above* $\hat{p}_l$ are characterised by payoffs strictly greater than the average, so that the relative frequency will decrease over time, and *vice versa.* Overall, then, frequencies are such that they will indeed *converge* to the steady state frequency $\hat{p}_l$. Obviously, local asymptotic stability implies Liapunov stability. *Global* asymptotical stability would hold when the basin of attraction of a steady state coincides with the whole region on which variables exist; that is there would exist only one local stable steady state. However, we shall not make use of this concept in the remainder of the Chapter. In the following sections, though, we shall indeed observe that the distinction between *Liapunov* and *asymptotical* stability is a key one for our analysis.

4.3.3 GPS stable steady states in the Prisoner's Dilemma with Normative Expectations

In what follows, I illustrate how the concept of GPS stable steady state can be used to test whether the first type of solutions reported in section 4.2.1, i.e. *inefficient* equilibria in the substitute strategy case, can be GPS replicator stable steady states. Notice that such a static equilibrium is certainly a trivial solution to the system (4.21). What needs to be checked is whether this steady state is stable. In order to do this, we first have to compute the average payoff of the population, which is made easier by the assumption that beliefs are consistent with the steady state strategy. The payoff of a generic *i*-player who is playing that equilibrium strategy is thus $V_i(\hat{p}_i;\hat{p}_i,\hat{p}_j)$. Therefore, the average player in population *i* will not experience any resentment, as her behaviour coincides with that of the bulk of the population: $m_i(\hat{p}_i;\hat{p}_i,\hat{p}_j)=0$.

---

[4] The generalisation of this condition for the n-dimension case would be as follows:
$\exists\ \omega>0$ s.t. $\forall\ \sigma_l \in \|\sigma_l - \hat{p}_l\| < \omega$, $\forall\ k=1..n$, $\{V(\sigma_k^l, \hat{p}_{-k}^l;\hat{p}_l)-\overline{V}(\hat{p}_l)\} \le 0$

Hence, her comprehensive payoff boils down to her material payoff, which is just symmetric to expression (4.5) above.

As for payoffs from 'deviant' behaviour, this, once again, varies in relation with whether we consider strategies 'above' or 'below' the average play level. Consider first the case of $\sigma_i > \hat{p}_i$. Here, the analysis is made easier by the shape of the resentment function: since super-erogatory actions are not rewarded with greater social approval, then the agent cannot gain any extra other-regarding utility from this type of action, thus the comparison between average payoff depends only upon the material component. But clearly the deviant agent gains an inferior payoff than the average, since Defection is the dominant strategy of the stage game. As a consequence, the density of any mixed strategy above the equilibrium level $\hat{p}_i$ is bound to decrease. Slightly more complex is the case of $\sigma_i < \hat{p}_i$, as now other-regarding utility does enter into play. However, the computation of comprehensive utility for the deviant agent in this case, shows that the sign of derivative (4.9) is determined by condition (4.4), which was the same underlying the static equilibrium concept. This implies that for all $\sigma_i < \hat{p}_i$ their frequency of play between deviant players is bound to increase (decrease) provided that $p_j < \overline{p}_j$. But this is indeed the case in regions surrounding the equilibrium, by construction of equilibria of type E_1.

Figure 4.6 shows the phase diagram of this case, drawn on the grounds of the foregoing analysis. Notice that the directions of the arrows signal the tendency of change of strategies within the sub-population of deviant agents. The result is clearly the *local* stability of the steady state coinciding with the static GPS equilibrium. The intuition is that there exists a tendency for deviant players to conform to the general behaviour of the majority of the population. Co-operating with higher probability than average is clearly inefficient as no gain is reaped. But also playing Defection with higher probability than average is not optimal, as the resentment induced in other-regarding utility outstrips the gain in material utility. Therefore, deviant behaviour will converge to average behaviour.

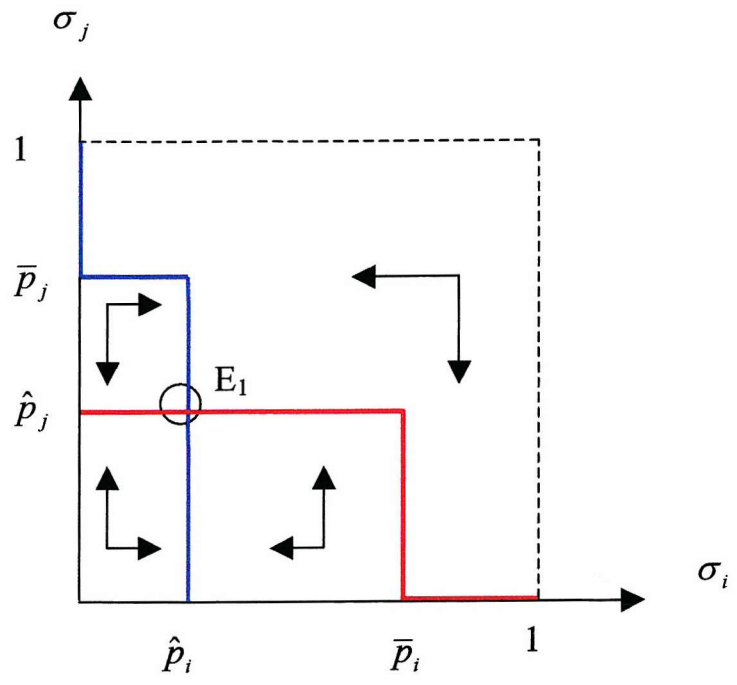

Figure 4.6

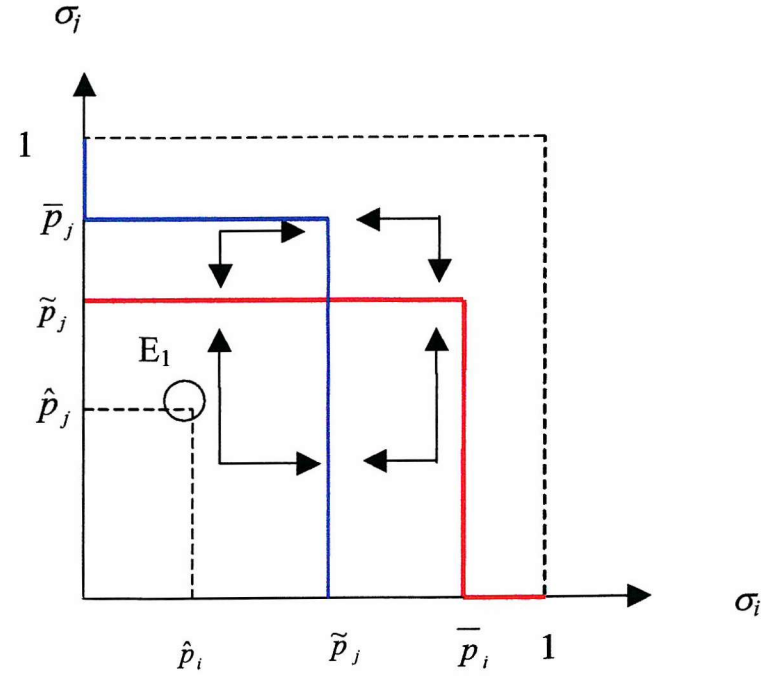

Figure 4.7

It is worth noticing that the condition determining the local stability of this steady state is the same as that which ensures that this is a Nash Psychological equilibrium of the game. This is not surprising, as the coherence between individual and average behaviour that we had imposed for the static concept of equilibrium (4.3) is clearly reminiscent of a dynamic notion of convergence. Moreover, the relationship between static Nash Psychological equilibria and stable GPS replicator steady states seems analogous to that between Nash equilibria and stable replicator steady states (see Weibull (1995); Fudenberg and Levine (1998)). In fact, since expectations are bound to be consistent with the equilibrium, the other-regarding component of utility will not be relevant in the comparisons between the payoffs, so that these can be carried out in terms of standard self-regarding utility functions. Though this appears a general result, a formal proof will not be provided here.

4.3.4 Deviations with Off-Equilibrium-Consistent Beliefs: The VK Replicator Steady State

The dynamic notion of stable dynamical equilibrium put forward in the previous section was based on the idea that deviant agents have beliefs consistent with the strategies played in the static equilibrium. This is tantamount to assuming that, whereas some deviant agents are performing a different behaviour from that carried out in equilibrium, the bulk of the population is already performing the steady state behaviour and this is common knowledge to deviants as well. There seems to be some ground to argue that such a concept of dynamical equilibrium actually requires *too little*, in that only the tendency of some negligible-size cohorts of agents to converge to the equilibrium is investigated, neglecting the question of whether there is the tendency for the *whole* population to converge, at least when starting within a suitably defined neighbourhood of the equilibrium. In other words, the GPS replicator steady state only studies the stability with respect to mutations by *0-measure* sub-sets of agents, but it does not deal with mutations of sets of agents with positive measure, thus falling short of some of the properties that a dynamical concept would be required to fulfil.

These considerations echo those put forward by Van Kolpin with regard to the original paper of GPS (Van Kolpin (1992)). In fact, some of the refinements put

forward by GPS, such as those of trembling-hand perfect equilibria, though not still dynamical in a strict sense, imply the study of optimal behaviour *outside* the equilibrium. Then, so Van Kolpin argues, beliefs should be designed to be consistent with the *actual* average play, rather than assuming consistency with the would-be equilibrium. This makes the analysis of behaviour probably more complicated, but surely more coherent with its own premises.

Building on these considerations, I shall propose a refinement of the previous concept of GPS stable steady state, which allows for the fact of significant deviations from the steady state behaviour, and beliefs that are built consistently with such deviations. On more practical grounds, this approach implies studying the rule of motion of deviant strategy when the average play differs from the steady state, *and* beliefs are consistent with such averages. Moreover, a similar distinction to that between stability in the Liapunov sense and in the local asymptotic sense that was put forward in relation to the GPS steady state, will also be proposed here.

A VK replicator steady state $\hat{p}_l$ is *stable in the sense of Liapunov* if, besides satisfying (4.21) and (4.20ii), it also fulfils the following condition:

$$\begin{aligned} &\exists\ \omega>0 \ \text{ s.t. } \ \forall\ \sigma_l \in \left|\sigma_l-\hat{p}_l\right|<\omega \ \textit{ and } \forall\ \tilde{p}_l \in \left|\tilde{p}_l-\hat{p}_l\right|<\omega\ , \\ &\left\{V(\sigma_l;\tilde{p}_l)\text{-}\overline{V}(\tilde{p}_l)\right\}(\sigma_l-\hat{p}_l)(\sigma_l-\tilde{p}_l)\le 0 \end{aligned} \tag{4.24}$$

Notice that this condition applies to the two-strategy case[5]. The main difference with respect to (4.22) is that the beliefs of deviant agents are now consistent with some average play $\tilde{p}_l$ lying in a neighbourhood of the steady state $\hat{p}_l$, rather than being coherent with $\hat{p}_l$ itself as in the GPS case. Local asymptotic stability requires the main inequality to hold strictly:

A VK replicator steady state $\hat{p}_l$ is *locally asymptotically stable* if, besides satisfying (4.21) and (4.20ii), it also fulfils the following condition:

$$\begin{aligned} &\exists\ \omega>0 \ \text{ s.t. } \ \forall\ \sigma_l \in \left|\sigma_l-\hat{p}_l\right|<\omega \ \textit{ and } \forall\ \tilde{p}_l \in \left|\tilde{p}_l-\hat{p}_l\right|<\omega\ , \\ &\left\{V(\sigma_l;\tilde{p}_l)\text{-}\overline{V}(\tilde{p}_l)\right\}(\sigma_l-\hat{p}_l)(\sigma_l-\tilde{p}_l)<0 \end{aligned} \tag{4.25}$$

---

[5] The generalisation of this condition for the n-strategy case would be as follows:
$\exists\ \omega>0$ s.t. $\forall\ \sigma_l \in \left\|\sigma_l-\hat{p}_l\right\|<\omega$, $\forall\ k=1..n,\ \left\{V\left(\sigma_k^l,\tilde{p}_{-k}^l;\tilde{p}_l\right)\text{-}\overline{V}(\tilde{p}_l)\right\}\left(\sigma_k^l-\hat{p}_k^l\right)\left(\sigma_k^l-\tilde{p}_k^l\right)\le 0$

Applying this concept to the analysis of the Prisoner's Dilemma seen in the previous section does imply a substantial difference, as Figure 4.6 shows. In fact, the same reasoning developed to analyse the previous case, now implies that the system will tend to gravitate around the *actual* average play $\tilde{p}_l = (\tilde{p}_i, \tilde{p}_j)$, provided that $\tilde{p}_l < \overline{p}_l$, $l = i,j$. In other words, there is no tendency for the system to move away from the current position and reach the 'designated' steady state $(\hat{p}_i, \hat{p}_j)$. At the light of the definitions of stability just put forward, we can conclude that $(\hat{p}_i, \hat{p}_j)$ is *stable* in the *Liapunov* sense, but not in the local *asymptotical* sense: given a steady state, the system will not depart away from a neighbourhood of the steady state, but it will not converge toward it either.

The reason for this result is that every sub-set of deviant agents will find it convenient to abide by what the bulk of the population is already doing: those who are Co-operating with higher probability than the average do not gain any reward for this, thus they will find it worthwhile to decrease their level of Co-operation; those who Co-operate with smaller probability than the average, provided that the *j*-player population is expected to perform a not too high amount of Co-operation that elicits Co-operation to an *i*-player, will experience resentment for causing a loss in utility to the opponent with respect to what expected, and this will outstrip the gain in material utility. Then, they will be prompted to increase their probability of Co-operation.

Therefore, although the VK criterion does not rule out steady states as *unstable*, it qualifies their stability as a Liapunov one, thus it implies that the system will lack a tendency to move away from its current position. This appears to be a general characteristic of this version of the normative expectations theory, which also carries over to the other types of equilibria that we have found, i.e. *anti-reciprocal*, or *exploitative*, and *efficient* ones.

In my view, this poses a serious threat to normative expectation theory, at least in the particular version we are studying. The reason is that such a theory does not seem capable of offering a dynamical account of how social norms can have emerged, leading to the conclusion that if the social norm was not already established, the system would have not tended to bring it about. This is suggestive of the lack of

incentives at the individual level to drive the system toward the emergence of social norms different from those already into place. It only stresses the existence of incentives to *uphold* a norm once this is in place, but it does not provide an account of how and why this norm has emerged. In this sense, in my view, it leaves unexplained most of what needs to be explained. Suppose, for instance, we wanted to offer an account of norms requiring Co-operation with probability (nearly equal to) one by (nearly) all members of the community. As a matter of fact, there would be no justification within this theory for this norm unless it was already established. The intuition of why this is so is very simple: since super-erogatory behaviour is not rewarded, no agent will have incentive to co-operate more frequently than what is currently the average behaviour, thus the system cannot progress towards universal Co-operation.

To be sure, assuming that the norm was already in place *for historic reasons* would simply beg the question, since it would merely shift the explanation one step backward. The problem would then become how to make clear which elements *in the past* have led to the establishment of the norm. It should be noticed that there is a clear difference with respect to evolutionary theory on this point. The latter does rely on history to explain the emergence of an outcome, but it is capable of making, at least in principle, testable predictions on this point. In fact, when saying that the final outcome depends on where the system was situated at the beginning of 'history', the evolutionary theorist makes a claim verifiable on empirical grounds, which can be contradicted by factual evidence. For instance, one could put forward the prediction that if a society started from a point belonging to a certain basin of attraction of a steady state, then the system would be bound to converge to the related steady state.

In the present model this is not possible, or it is possible only to a very limited extent, since the model shows a peculiar tendency for all agents to conform to the current average behaviour. Such a propensity is clearly consistent with the considerations put forward in section 3.5 about the 'conformative' character of normative expectations: these act as a powerful tool to attract agents to imitate other agents, and this aspect leads to the absence of proper 'evolution' in the present setting. The underlying reason for this is the 'empirical' nature of such expectations, which are

only related to past history but not to some 'causal' justification of the current situation, as argued in section 3.5.2. The result is that no clear-cut prediction is possible in this model, as the system will not evolve away from its current position, provided that it is situated in a region defining static equilibria, i.e. those associated with the sets E_1, E_2 and E_3 in the above sections. With the model of the next section I hope to be able to show how a different account of individual motivations can help to overcome some of these shortcomings.

4.4 Other-Regarding Motivations Grounded on a Shared View of Morality

4.4.1 The Relationship Between Normative Expectations and The Moral Point of View

In the foregoing section I have illustrated some aspects of normative expectations theory. The purpose of that analysis was twofold: one was mainly methodological, and consisted of underpinning a dynamical analysis of how norms get established in a society. On the other hand, I wanted to make clear some unsatisfactory features that, in my view, characterise that theory. In particular, my criticism relied on the distinction between 'causal' and 'empirical' expectations put forward in section 3.5, and was based on the idea that only the former type of expectation can have an effective motivational force for individuals. In this section I would like to contrast the results previously obtained with those that would be reached if individual behaviour followed the model of conditional compliance with morality that has been developed in section 2.5. In this section I will expand on the relationship between the two approaches, and try to clarify possible criticism in the application of the latter model.

The first criticism is that my model, by taking for granted a notion of morality, fails to comply with the 'reductionist' approach that characterises most of the contributions in the literature on social norms (Sugden (1998a); Binmore (1998)). Simply stated, the reductionist approach aims to ground morality on some naturalistic feature of reality. On Sugden's view, being willing to avoid the resentment of other members of a community can be thought of as a general characteristic of human

behaviour, and thus an account of morality grounded on this idea can be said to satisfy the reductionist claim. Binmore's account of morality can also be said to lie within this realm, since it emphasises how morality ultimately stems from individual behaviour, where individuals are depicted as boundedly rational agents who slowly adjust to locally optimal behaviour. Therefore, so the argument would go, by supposing that some ideas of morality is already established, I would fail to account for it in terms of some 'naturalistic' features of individual human behaviour. Rather than *accounting for* morality in terms of its naturalistic, or evolutionistic, features, I would *adopt* morality to study its motivational force on individuals and its impact on social outcomes by means of the aggregation of the corresponding individual actions.

However, I believe that the structure of my argument does not substantially differ from that of theorists in the naturalistic tradition. What will be relevant in my account, in fact, is not so much *which* type of notion of morality is established, but that agents perceive the compliance with it as a relevant prompt to action in their system of motivations. In other words, the willingness to avoid the breaching of some shared notion of morality, whatever this may be, may be deemed as a 'natural' characteristic of human motivations in the same way as the willingness to avoid others' resentment. That is, both can be seen as inborn traits of human beings' behaviour, so that even my account could be thought of as consistent with a naturalistic approach. In other words, the real difference between the approach of Sugden and Binmore and that pursued here is that the present one opens up the possibility of motivations that respond to social or group characteristics rather than purely individual characteristics. It is the fact that individuals are concerned with some notion of public interest rather than individual resentment that is distinctive, but this social aspect of human beings can certainly be viewed as 'natural'.

Moreover, the resentment hypothesis and the conditional compliance with morality hypothesis, rather than being conflicting accounts of individual behaviour, are probably best seen as complementary to each other. In particular, let me suppose that morality is associated with some clear-cut ideas of a public interest. Then, the willingness to comply with it may well be seen as accompanied and strengthened by the resentment that people would experience when failing to comply with it. For it is

easy to envisage that an individual disrupting what is commonly viewed as public interest would trigger the resentment of other members of the community, so that the compliance with norms embodying some notions of public interest would come to be associated with normative expectations. Public interest may instead offer the 'causal' feature that is necessary to make resentment an effective prompt to action. Hence, individuals would attach a disutility to the perception of having gone against the normative expectations of the community, but only *insofar as* they perceive that by doing so they are violating the public interest. In other words, according to the model of this section, normative expectations come to have a binding role on individual reasons to action when there exists some notion of public interest that justifies conformity with those expectations.

However, although resentment is likely to accompany the breaching of a norm embodying public interest, one should note that this relationship is an ancillary one. That is, the motivation to abide by the public interest is independent from the normative expectations that may come with it. This is evident if one thinks at the numerous cases in which people act in accordance with some clearly recognisable ideas of public interest even when there are no expectations of any kind on them to do so. Collective action groups such as environmentalists, or civil/human rights activists, or campaigners for political issues are all clear examples of this. Although clearly there is no normative expectations on them from the society they live in, they are motivated to act in the way they act by the belief that their action endorses some ideas of public interest.

In the following section I will employ the particular notion of morality employed in section 2.6, i.e. the Nash social welfare function. This can be taken as a representation of a contractarian account of justice, whose properties seem to satisfy the requirements of impersonality depicted in section 1.4. In the conclusive section of this chapter I will also expand on the results that the adoption of different particular conceptions of the public interest would have caused.

4.4.2 Specification of the Model for a Continuum of Populations

The model of individual choice developed in section 2.5 was initially set up to fit with a generic two-person relationship. Consequently, we now need to adapt it to the evolutionary setting in which the present analysis is being conducted. In particular, we have to take into account two different aspects. The first concerns the consideration of a 'continuum' of agents, where the measure of each agent is negligible with respect to that of the entire population. The second relates to the presence of two distinct populations of agents, whose members are drawn at random and matched to play a stage game. Both aspects will be dealt with by developing the approach previously developed in this Chapter. In particular, the agent will consider the *average behaviour* in either population as the relevant quantity against which to gauge hers and her opponent's expected degree of compliance to the normative principle. In particular, it is assumed, as before, that beliefs of first and second order are *consistent* with such average plays.

The model developed in section 2.5.3 was grounded on the idea that the agent estimates her own and her opponent's compliance with the shared normative principle, given her first and second order expectations on each other behaviour. Let us consider how this idea adapts to this new framework. First the agent's own estimated compliance with the normative principle, given her first order belief on her opponent's behaviour, is given by the following expression:

$$f_i\left(\sigma_i, \beta_i^1(p_j)\right) = \frac{T(\sigma_i, p_j) - T^{MAX}(p_j)}{T^{MAX}(p_j) - T^{MIN}(p_j)} \tag{4.26}$$

Recall from above notation that $\beta_i^1(p_j)$ represents the first order belief *consistent* with p_j. i.e. the i-player attaches probability one to the fact that her opponent will follow the mixed strategy associated with p_j. As in section 2.5.2, $T^{MAX}(p_j) = \arg\max_{\Sigma_i} T(\sigma_i, p_j)$ and $T^{MIN}(p_j) = \arg\min_{\Sigma_i} T(\sigma_i, p_j)$. That is, $T^{MAX}(p_j)$ and $T^{MIN}(p_j)$ represent, respectively, the maximum and the minimum value that the normative function can assume, given i's first order belief on player j's behaviour. Therefore, if $T^{MAX}(p_j)$ is obtained, then

agent *i* is maximising the normative function given her first order belief on player *j*'s behaviour. Conversely, if $T^{MIN}\left(b_i^1\right)$ obtains, then the *i*-player is minimising the normative function. $T\left(\sigma_i, b_i^1\right)$ is instead the value of the normative function corresponding to *i*'s actual choice σ_i. In the remainder of the chapter, I shall express (4.27) as a direct function of p_j, thus omitting the operator β of consistency in beliefs. The interpretation of (4.26) is the same as that given in 2.5.2: it is an index varying from –1 to 0, which is higher the closer *i*-player's action is to the maximisation of the normative function *T*.

The expected compliance of player *j* to the normative criterion can be derived along the same lines. In particular, the *i*-player will take into account her *second* order beliefs, which are consistent with the *average play of her own population*, rather than on her own behaviour. In fact, the *i*-player is aware that her opponent will base his expectations on the average play of the population from which *i* is randomly drawn to play, rather than on *i*'s actual behaviour. Therefore, the expected compliance of *j* to the normative principle will be as follows:

$$\tilde{f}_j\left(\beta^1\left(p_j\right), \beta^2\left(p_i\right)\right) = \frac{T\left(p_i, p_j\right) - T^{MAX}\left(p_i\right)}{T^{MAX}\left(p_i\right) - T^{MIN}\left(p_i\right)} \qquad (4.27)$$

where $T^{MAX}\left(p_i\right) = \arg\max_{\Sigma_j} T\left(p_i, \sigma_j\right)$ and $T^{MIN}\left(p_i\right) = \arg\min_{\Sigma_j} T\left(p_i, \sigma_j\right)$ represent respectively the expected value that the normative function takes when a generic *j*-player maximises or minimises it, given the second order belief of player *i*. In other words, those functions indicate the maximum and minimum values that a *j*-player can attribute to the normative function, given the belief he has about an *i*-player action, as computed by *i* herself. Alike the twin functions (4.24), a value of (4.27) equal to 0 (-1) indicates that a *j*-player is expected by *i* to comply with the normative function at its maximum (minimum) degree. Notice that such a function is independent from *i*'s own action, as all the relevant beliefs are based on the average behaviour of any population.

The comprehensive utility function can thus be introduced:

$$V_i\left(\sigma_i; p_i, p_j\right) = U_i\left(\sigma_i, p_j\right) + \lambda\left[1 + \tilde{f}_j\left(p_j; p_i\right)\right]\left[1 + f_i\left(\sigma_i; p_j\right)\right] \qquad (4.28)$$

The intuition is the same as that provided in Chapter 2: overall utility is given by the sum of material and other-regarding utility, with λ acting as a weight on the two sources of utility. In the second component, the opponent's esteemed compliance with the normative criterion still acts as a 'marginal incentive' for one's own compliance, but, given the new setting, this time this is independent on *i*'s own action.

4.4.3 The Nash Social Welfare Function

The properties of the Nash social welfare function as a normative criterion have already been emphasised by many authors (see Harsanyi (1977); Brock (1979)), and it is well known that it can be held to represent a contractarian account of social justice. One of the main characteristics of the Nash welfare function is its dependence on the *status quo* of the bargaining process. Since in many social situations it is not thoroughly clear where this should be located, its choice represents a relevant issue[6]. In the present section, I will assume that the status quo is given by the standard Nash solution of the game, i.e. the outcome in which all agents defect. As we shall observe, this is a key determinant of the results we shall obtain. However, the utilitarian case in the following section could be seen as a version of the Nash function under a different choice of the status quo from the one made here. Finally, as well as in earlier analysis, I will focus on intermediate values of λ such that no action is strategically dominant for each player.

Recalling the expression of the Nash social welfare function given in equation (2.19), which in the remainder of this section will be denoted as $N(\sigma_i, \sigma_j)$, the matrix of the 'ideal game' is as follows:

	Co-operation	Defection
Co-operation	$(\gamma-\alpha)^2$	$(\delta-\alpha)(\beta-\alpha)$
Defection	$(\delta-\alpha)(\beta-\alpha)$	0

Figure 4.8

[6] In fact, one of the main distinctive features of theories of distributive justice lies in how the status quo is determined: for different accounts of the status quo, see Buchanan (1975), Nozick (1975), Gauthier (1986), and, of course, Rawls (1971).

What is apparent is that mutual Co-operation is the outcome maximising the normative criterion. What is perhaps less obvious is that the two outcomes in which one player Co-operates as the other Defects receive an even lower evaluation than that of mutual Defection. The reason for this hinges upon the choice of the status quo that has been made: since the player who Co-operates receives a lower material payoff than would otherwise be gained in the status quo, then the Nash criterion assigns a negative evaluation to this outcome. The underlying reason for this result is related to the feature of the contractarian criterion as guaranteeing the respect of some minimal 'rights' to the agents involved in the interaction. In the present case, such rights are represented by means of the payoff that the player gains in the status quo. In other words, according to a contractarian point of view, a payoff going below the status quo level can be interpreted as a violation of some fundamental rights of the player, which explains the negative value assigned to this outcome by the normative criterion. For the same reasons, the Nash criterion assigns an evaluation of zero to the case of mutual defection.

4.4.4 Equilibria in a Prisoner's Dilemma

First, let us deal with static equilibria, i.e. those fulfilling (4.3). It is easy to show how the outcomes associated with Co-operation with probability one by either population and Defection with probability one in either population are both equilibria of the game. Let us look at the agent's own degree of compliance with the normative criterion $N(\sigma_i, \sigma_j)$, given expectations consistent with the average play p_j in the other population. The main point of this analysis is that the action that satisfies the normative criterion to the largest extent *changes* depending on whether p_j is above or below a certain threshold level. The reason is that if the j-player is expected to Co-operate with relatively high probability, then Co-operation is the action maximising $N(\sigma_i, \sigma_j)$ under this expectation. This is clear if one considers that mutual Co-operation is the Pareto-efficient outcome of this game. However, if the opponent is expected to Co-operate with relatively low probability, then Co-operating would make an 'exploitative' equilibrium in which the i-player Co-operates and the j-player Defects

very likely. As suggested above, this is the outcome that receives the worst evaluation from the standpoint of the normative criterion, the reason being that such an outcome is contrary to the main tenets of a contractarian criterion that some basic individual rights should be preserved from the status quo level, or, in other words, that contractarianism awards those allocations that *mutually advantage* the parties of the 'contract'. As a result, the *i*-player should now see Defection as the action best fulfilling $N(\sigma_i, \sigma_j)$. In formal terms:

$$\arg\max_{\Sigma_i} N(\sigma_i, p_j) = \begin{cases} C & \text{if } p_j \geq \overline{p} \\ D & \text{otherwise} \end{cases} \tag{4.29}$$

where

$$\overline{p} = \frac{(\beta - \alpha)(\alpha - \delta)}{(\gamma - \alpha)^2 + 2(\alpha - \delta)(\beta - \alpha)} \tag{4.30}$$

Notice that such a threshold level has not been indexed, because, given the symmetry of the game, the same threshold would apply to the *i*-player population with respect to the *j*-player's decisions.

I believe two properties of this result are worth emphasising. First, although the general criterion of normative assessment is fixed in advance, its practical implications are variable over time, both in the sense that the practical motivational force that norms have on individuals depends on the *general* degree of compliance with it, but also, more importantly, in the sense that there could be 'moral regime switches' when average behaviour goes through this threshold in one or in the other direction.

Second, the analysis emphasises the different perspective that an agent should take when assessing the situation from her own standpoint and from the normative standpoint. In fact, in the 'exploitative' case, it is the *exploited* agent who, as well as the other, attaches a negative value to the normative function. In other words, it is the very agent who is performing a self-sacrificing action, which actually brings about extra benefit to her counterpart at the expense of her own utility, who 'ought' to comprehend the negative evaluation of this action from the normative point of view. In this sense, the normative criterion acts as a safeguard for the very agents who are being exploited, *which acts upon their own system of motivations.*

Let us now analyse the implications of the existence of such a threshold on the type of feasible equilibria in the static point of view. First, let us check whether the outcome in which all agents Co-operate, or Co-operate with high probability, can be sustained as a Nash Psychological equilibrium of the game. The answer is indeed positive. First, let us suppose that $p_j \geq \overline{p}$, so that the best action in terms of the normative criterion is to Co-operate, which of course pulls in opposite direction than the self-interested action, which would be to Defect. On the grounds of what probably is by now a customary analysis, the final decision of the agent crucially hinges upon her value of λ. If this is sufficiently large, than the overall best strategy for the individual is indeed to Co-operate. In particular, the agent's own compliance to $N(\sigma_i, \sigma_j)$ boils down to the following simple expression[7]:

$$f_i\left(\sigma_i, \beta_i^1\left(p_j \middle| p_j \geq \overline{p}\right)\right) = -(1-\sigma_i) \tag{4.31}$$

This tells us that if an *i*-player wants to maximise her compliance with $N(\sigma_i, \sigma_j)$, provided that the percentage of *j*-players who Co-operate exceeds the threshold, then she has to Co-operate as well. Now we have to evaluate the esteemed compliance for a generic *j*-player with the normative criterion. As pointed out above, this is computed in terms of the expected *average* compliance to $N(\sigma_i, \sigma_j)$ of the *i*-players population. Let us suppose that $p_i \geq \overline{p}$. Then, a *j*-player would have the same attitude toward the judgement of the action that best adheres with $N(\sigma_i, \sigma_j)$ as that illustrated above for an *i*-player. That is, he would associate the best action in terms of the normative criterion with Co-operating, and the worst to Defecting. Therefore, the esteemed compliance of a *j*-player from the point of view of an *i*-player is given by the following expression, which is symmetric to (4.31):

$$\tilde{f}_j\left(\beta^1\left(p_j\right), \beta^2\left(p_i \middle| p_i \geq \overline{p}\right)\right) = -(1-p_j) \tag{4.32}$$

As a result, the comprehensive utility function takes on the following expression:

$$V_i\left(\sigma_i; p_i, p_j\right) = U_i\left(\sigma_i, p_j\right) + \lambda_i p_j \sigma_i \tag{4.33}$$

---

[7] The fact that the exact magnitude of the values of $T^{MAX}\left(p_j\right)$ and $T^{MIN}\left(p_j\right)$ do not appear in this expression depends on the presence of only two pure strategies in the agent's actions set. If the number of pure strategies was greater than two, then these values could not be simplifies, and in general a more complex expression would obtain.

As usual, the final result depends on the constellation of parameters that is chosen. However, a few conditions need to be analysed, and they turn out to be quite intuitive. First, let us suppose that:

$$\lambda > \beta - \gamma \tag{4.34}$$

The meaning for this is quite straightforward: $\beta-\gamma$ represents the opportunity cost, in terms of material utility, of not Defecting when the counterpart is Co-operating with probability one. Thus, (4.34) requires the marginal benefit of other-regarding utility to be greater than the material opportunity cost of performing the action maximising the normative criterion.

Under this condition, it is then straightforward to show that Co-operate is the best action for the *i*-player in terms of her comprehensive utility. In fact, differentiating (4.33) with respect to σ_i, one finds that an *i*-player best reply function is as follows:

$$\arg\max_{\Sigma_i} V(\sigma_i; p_i, p_j) = \begin{cases} C & if\ p_j \geq \overline{\overline{p}} \\ any & if\ p_j = \overline{\overline{p}} \\ D & otherwise \end{cases} \tag{4.35}$$

where:

$$\overline{\overline{p}} = \frac{\alpha - \delta}{\lambda - \eta} \tag{4.36}$$

Let us assume that p_j is relatively high, so that it outstrips both the relevant thresholds: namely, $p_j \geq \max\{\overline{p}, \overline{\overline{p}}\}$; then, we can be sure that both the relevant conditions in terms of *j*-player's behaviour are fulfilled, thus ensuring that Co-operation is the best strategy for a generic *i*-player.

Let us now assume that a symmetric condition to that on p_j held for *i*-players as well. That is:

$$p_l \geq \max(\overline{p}, \overline{\overline{p}}) \quad for\ l = i,j \tag{4.37}$$

Given the symmetry of the game, under symmetric conditions on *i*-player' behaviour, Co-operation will be the best strategy for a *j*-player as well. Therefore, the outcome in which $(\hat{p}_i = 1, \hat{p}_j = 1)$ - namely, the outcome in which players in either population Co-operate with probability one – is a Nash Psychological equilibrium of the game.

The intuition for this result is also quite clear: provided that other-regarding utility gives the agent sufficient motivational force within comprehensive utility, which is ensured by condition (4.34), then a deviation from the situation in which all players of either population Co-operate and are expected to Co-operate with probability one determines a loss in other-regarding utility that is not compensated by the corresponding gain in material utility. This occurs because compliance by a generic *j*-player is expected to be relatively high, i.e. it is greater than the threshold level (4.36), thus spurring the incentive to Co-operate by the *i*-player herself to the maximum possible degree in (4.33).

Unlike the normative expectation model, in which various continuums of mixed strategy equilibria emerged, in the present case the shape of the individual best-reply function (4.33) and the consistency condition between individual and aggregate play embedded in (4.3), rule out almost all mixed strategy equilibria. Only one of these survives in the present setting, which is $\left(\hat{p}_i = \overline{\overline{p}}, \hat{p}_j = \overline{\overline{p}}\right)$. It holds under the conditions on parameters mentioned above, plus a new one: $\overline{\overline{p}} \geq \overline{p}$. In fact, when the opposite player is performing $\hat{p}_j = \overline{\overline{p}}$, then player *i* is indifferent between Co-operating and Defecting, thus in particular she can play $\hat{p}_i = \overline{\overline{p}}$. Given the symmetry of the game, this can thus be sustained as an equilibrium.

What happens if the condition (4.34) does not hold? In this case, the structure of the best-reply function is reversed with respect to (4.35):

$$\arg\max_{\Sigma_i} V\left(\sigma_i; p_i, p_j\right) = \begin{cases} C & if\, p_j \leq \overline{\overline{p}} \\ any & if\ p_j = \overline{\overline{p}} \\ D & otherwise \end{cases} \tag{4.38}$$

Therefore, $\left(\hat{p}_i = 1, \hat{p}_j = 1\right)$ cannot be an equilibrium any longer. The reason is that now the weight attached to the other-regarding utility is not high enough to make inconvenient to an *i*-player deviating from Co-operation when any other agent is Co-operating. That is, material utility now compensates the lesser other-regarding utility in the case of a deviation. However, provided that, as above, $\overline{\overline{p}} \geq \overline{p}$, the structure of incentives is such that in the interval $p_j \in \left(\overline{p}, \overline{\overline{p}}\right)$ player *i* is willing to Co-operate. In

this region, the lesser degree of Co-operation by a *j*-player makes the deviation less attractive in material terms, thus the agent opts for Co-operation. This makes it possible that the same mixed strategy equilibrium as the previous case obtains: $\left(\hat{p}_i = \overline{\overline{p}}, \hat{p}_j = \overline{\overline{p}}\right)$.

It remains to be seen whether other equilibria are possible in the regions where at least one of the two populations average play is below $\overline{p}$, which would bring about an exploitative type of equilibrium (see section 4.2). This will then help a comparison with normative expectations theory for what concerns the emergence of this somehow puzzling outcome. It can be shown that in the present setting this cannot be an equilibrium of the game. Suppose in particular that $p_i \geq \overline{p}$ and $p_j < \overline{p}$ and let us study the incentives that an *i*-player has in complying with this situation. Recalling (4.29), the best action in terms of the normative criterion is indeed to Defect. Therefore, *i*'s own compliance with the normative criterion is as follows:

$$f_i\left(\sigma_i, \beta_i^1\left(p_j \middle| p_j < \overline{p}_j\right)\right) = -\sigma_i \tag{4.39}$$

That is, *i*'s compliance with N is maximised when she sets σ_i=0. Instead, given *i*'s high degree of Co-operation, the best action for a *j*-player in terms of the normative criterion would be to Co-operate. Hence, the expected degree of compliance of a generic *j*-player is as in (4.32). Combining these results together, we have that the comprehensive utility function is now:

$$V_i\left(\sigma_i; p_i, p_j\right) = U_i\left(\sigma_i, p_j\right) + \lambda_i p_j\left(1-\sigma_i\right) \tag{4.40}$$

Differentiating this expression with respect to σ_i, one can find that in this case the best action is always to Defect. The intuition is quite clear: both other-regarding and material utility prescribe Defection as the most preferred action, thus this remains the most preferred outcome when integrating the two perspectives together. Hence, relying on the interpretation of the contractarian criterion as guarantying the safeguard of some 'minimal' individual rights, or protecting the idea of 'mutual advantage', once it is assumed that this is internalised into the motivational system by every agent in the community, we can be sure that exploitative equilibria are ruled out from the game.

The symmetric case to that just analysed, in which $p_i < \overline{p}$ and $p_j \geq \overline{p}$, would probably be more interesting from the strategic point of view for an *i*-player, since

now other-regarding and material utility would be maximised by different actions: self-interest would prescribe Defection (as usual), whereas concern for the normative criterion would elicit Co-operation, given the high degree of Co-operation from the other population. However, we shall omit this analysis as irrelevant, since, given the symmetry of the game, a *j*-player would always find it optimal to Defect in this situation, thus causing the impossibility of reaching any equilibria in this region.

The last case is that in which $p_i < \overline{p}$ and $p_j < \overline{p}$. It is clear, on the grounds of the above analysis, that both players will find it optimal to Defect in this setting; thus the outcome characterised by Defecting with probability 1 by either population is an equilibrium, and it is the only one in this region due to the consistency condition implied by (4.3).

4.4.5 The Stability of Equilibria

In this section, I will employ the concepts of dynamically stable equilibria put forward in sections 4.3.2 and 4.3.4 to check whether these are also stable from a dynamical point of view. The main result is that the pure strategy equilibria can indeed be proven to be stable, whereas the mixed strategy equilibrium is instead unstable.

Let us first analyse the Co-operative equilibrium: $(\hat{p}_i = 1, \hat{p}_j = 1)$, which held under condition (4.34) and (4.37). In order to test whether this is stable to any of the concepts put forward earlier, we need to compute the average payoff for a generic *i*-player under the fully Co-operative equilibrium. Recalling the expression of the comprehensive utility function (4.33), and substituting the relevant average play for p_i and p_j, we obtain the following expression for average payoff in the *i*-player population:

$$\overline{V}_i(\hat{p}_i, \hat{p}_j) = \gamma + \lambda_i \tag{4.41}$$

Now let us derive the payoff for a deviation from some *i*-player. Let us first employ the GPS criterion for stability, whose main assumption, as reported in (4.22) and (4.23), is that beliefs of deviant agents are consistent with the steady state play $(\hat{p}_i = 1, \hat{p}_j = 1)$. A deviant's payoff would then become:

$$V_i(\sigma_i; \hat{p} = 1_i, \hat{p}_j = 1) = \sigma_i\gamma + (1-\sigma_i)\beta + \lambda_i\sigma_i \tag{4.42}$$

As expected, such a deviation brings about higher material utility and smaller other-regarding utility.

According to the replicator dynamics as expressed in (4.17), the percentage of deviant players using a generic strategy σ_i will increase if this brings about a higher payoff than the average. We have already discussed how this use of the replicator dynamics is not entirely satisfactory in that other-regarding utility is part of the 'success' that agents want to imitate, but I shall be content with the remarks put forward in section 4.3.1. After some calculations, the growth rate of players using σ_i is then equal to:

$$\frac{\dot{\vartheta}(\sigma_i)}{\vartheta(\sigma_i)} = V_i(\sigma_i) - \overline{V}_i = (1-\sigma_i)(\beta - \gamma - \lambda) \qquad (4.43)$$

This set of differential equations is obviously solved by setting σ_i=1, which confirms that $(\hat{p}_i = 1, \hat{p}_j = 1)$ is a GPS replicator steady state, according to (4.20) and (4.21). The question now becomes whether this is a stable steady state or not. The answer is indeed positive: it suffices to note that, holding condition (4.34), the sign of all the growth rates in (4.43) is always negative but for $\sigma_i = 1$, thus implying that the frequencies of all the strategies except $\sigma_i = 1$ will decrease over time. Hence, $(\hat{p}_i = 1, \hat{p}_j = 1)$ turns out to be stable in the *asymptotic* sense.

The same is true if the more stringent concept of stability, i.e. the VK criterion, is used. Recall from (4.24) and (4.25) that such a concept does not require expectations of deviant players be consistent with the equilibrium strategies; instead, expectations are supposed to be consistent with an average play arbitrarily close to the equilibrium. This enables us to study the tendency of the system to 'move toward' the equilibrium even when the average player performs a different strategy. Let us suppose that the beliefs of a generic deviant *i*-player are consistent with an average play that we denote with $(\tilde{p}_i, \tilde{p}_j)$. Furthermore, we require that a condition analogous to (4.37) is satisfied for the *current* average play:

$$\tilde{p}_l \geq \max(\overline{p}, \overline{\overline{p}}) \quad for\ l = i,j \qquad (4.44)$$

This ensures that we are considering a neighbourhood of the point $(\hat{p}_i = 1, \hat{p}_j = 1)$. Condition (4.44), as will become clear immediately, delimits the basin of attraction of this steady state. Given these constraints, we can determine the function of esteemed compliance with the normative criterion for a *j*-player and of the *i*-player herself, which derive from (4.31) and (4.32) by simply substituting the relevant average play $(\tilde{p}_i, \tilde{p}_j)$ for (p_i, p_j). After having replaced these expressions into the comprehensive utility function (4.33), and having computed average payoff, we find that the growth rate of a generic deviant strategy σ_i is equal to:

$$\frac{\dot{\vartheta}(\sigma_i)}{\vartheta(\sigma_i)} = U_i(\sigma_i, \tilde{p}_j) - U_i(\tilde{p}_i, \tilde{p}_j) + \lambda_i \tilde{p}_j (\sigma_i - \tilde{p}_i) \qquad (4.45)$$

Simplifying this expression gives us:

$$\frac{\dot{\vartheta}(\sigma_i)}{\vartheta(\sigma_i)} = (\sigma_i - \tilde{p}_i)\{(\lambda - \eta)\tilde{p}_j - (\alpha - \delta)\} \qquad (4.46)$$

The second term of (4.46) is always positive in the neighbourhood of $(\hat{p}_i = 1, \hat{p}_j = 1)$ by construction, whereas the sign of the first factor depends on whether σ_i is greater than the average play or not. Therefore, all the strategies σ_i that are greater than $\tilde{p}_i$ will tend to grow, whereas all the others are bound to decrease. This implies that, overall, the average play will grow over time, and will tend to 1. This ensures the stability of the equilibrium in the asymptotic sense with respect to the VK criterion.

The intuition for this result is quite simple: since we are in the region where a *j*-player Co-operates with relatively high probabilty, and the propensity to follow the other-regarding motive by a *i*-player is relatively high, then the best strategy for her is to Co-operate. As a result, the strategies that perform Co-operation with higher probability than the average will fare better than those that prescribe Co-operation with lower probability. This implies that, over time, players will imitate the most 'successful' behaviour, in terms of comprehensive utility, of the highly co-operative players, thus on average Co-operation will increase until it reaches the value of 1.

Similar analysis shows that the other pure strategy equilibrium, which prescribes Defection with probability one to either population, is asymptotically stable under

both criteria. It is then interesting to check whether the mixed strategy equilibrium turns out to be dynamically *unstable*, as this is what frequently occurs in the face of refinements of static equilibria through dynamical considerations. The result of this analysis is that while the GPS criterion is trivially satisfied for the mixed strategy equilibrium, thus posing some doubts on its suitability as a refining concept with respect to static equilibria, the VK criterion identifies the instability of such a steady state.

Let us briefly analyse the two criteria in turn. Under the GPS dynamical criterion, beliefs of deviant agents are required to be consistent with the equilibrium $\left(\hat{p}_i = \bar{\bar{p}}, \hat{p}_j = \bar{\bar{p}}\right)$. Let us assume that $\bar{\bar{p}} \geq \bar{p}$, under which this equilibrium held in the static sense under condition (4.34) (see section 4.4.4). Hence, if one computes the growth rate in the frequency of a generic σ_i, he will find that this is equal to:

$$\frac{\dot{\vartheta}(\sigma_i)}{\vartheta(\sigma_i)} = (\sigma_i - \bar{\bar{p}}_i)\{(\lambda - \eta)\bar{\bar{p}}_j - (\alpha - \delta)\} \tag{4.47}$$

Now, the second term is always equal to zero due to the definition of $\bar{\bar{p}}$, so that all growth rates are equal to zero. This result is due to the 'threshold' character of $\bar{\bar{p}}$, to which average play is equal in the equilibrium that we are testing. In fact, when average play of a *j*-player equals $\bar{\bar{p}}$, the *i*-player is indifferent between Co-operating and Defecting, thus any strategy will fare the same as any other and as, in particular, the *i*-player population average play itself. Hence, no tendency for the system to evolve will be displayed.

Conversely, the VK criterion does imply the instability of this mixed strategy equilibrium. Again, let us assume that players' beliefs are consistent with some $(\tilde{p}_i, \tilde{p}_j)$ that is arbitrarily close, but not necessarily coincident, with the static equilibrium $\left(\hat{p}_i = \bar{\bar{p}}, \hat{p}_j = \bar{\bar{p}}\right)$. In particular, suppose that the condition that $\bar{\bar{p}} \geq \bar{p}$ holds even in this case. Then, the evolution of strategies frequency will be driven by the following equation:

$$\frac{\dot{\vartheta}(\sigma_i)}{\vartheta(\sigma_i)} = (\sigma_i - \tilde{p}_i)\{(\lambda - \eta)\tilde{p}_j - (\alpha - \delta)\} \tag{4.48}$$

If we consider the section of the neighbourhood of $(\bar{\bar{p}}, \bar{\bar{p}})$ in which $\tilde{p}_j < \bar{\bar{p}}$ and $\tilde{p}_i < \bar{\bar{p}}$, we have that strategies σ_i such that $\sigma_i < \tilde{p}_i$ perform better than the average, thus the average will decrease over time and the system will *diverge* from the equilibrium. In fact, in this region the frequency of co-operative behaviour in the *j*-player population is below the threshold level that elicit co-operation. Consequently, a less co-operative behaviour than the average by an *i*-player will be awarded with higher comprehensive utility then the average, which accounts for the divergence of the system from the steady state. The same divergent behaviour can be observed in the section of the neighbourhood of $(\bar{\bar{p}}, \bar{\bar{p}})$ where $\tilde{p}_j > \bar{\bar{p}}$ and $\tilde{p}_i > \bar{\bar{p}}$. Here, the threshold is exceeded thus *more* co-operative behaviour fares better than the average, so that average play in the *i*-player population will be growing over time and moving away from the steady state. These considerations are represented in Figure 4.9, which depicts qualitatively the tendency of average play to evolve over time, depending on the initial position of average play $(\tilde{p}_i, \tilde{p}_j)$.

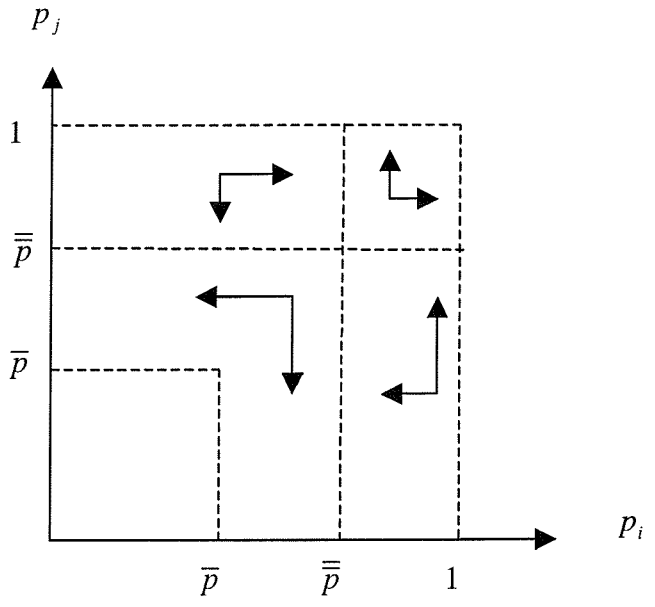

Figure 4.9

Figure 4.9 highlights that in some of the regions surrounding the steady state, the system is bounded away from it, if initial beliefs are consistent with average play belonging to those regions.

Therefore, the concept of stable VK steady state does prove itself to be a {useful or appropriate} refinement of the static equilibrium notion, and, by analogy to what occurs in 'standard' Game Theory, it rules out mixed strategy equilibria while confirming the stability of pure strategy equilibria. In fact, there is ground to suppose that between Nash Psychological equilibria and VK stable steady state there exists the same relationship as exists between standard Nash equilibria and stable replicator steady states, where the latter can be proven to be a refinement of the former. Although a formal proof of this result will not be attempted here, this seems to be an interesting development in this line of research.

4.5 Different Normative Criteria, Rights and Normative Expectations

With the analysis developed in the previous section, I hope that two points have been shown with sufficient clarity: the first is that if agents are concerned with the *substantive* content of their other-regarding actions, and this is shaped in accordance with the contractarian ideas that some minimal rights have to be attributed to each agent, then those 'counter-intuitive' exploitative equilibria that emerged within the normative expectations theory are ruled out. More generally, if agents are supposed to pay attention to the 'substantive' character of the normative principle to which they attach importance, then the 'herding' effect typical of the normative expectations theory also disappears, as the previous dynamical analysis has demonstrated. The other point that is worth mentioning is that 'dynamical' analysis of such equilibria is indeed a powerful tool in order to shed light on the cognitive and strategic structure underlying them. It shows that some equilibria should be ruled out as 'unstable', or stable only in the Liapunov sense, since the system itself would not manifest a tendency to evolve towards them unless that steady state was already in place. In this concluding section, I would like to comment more extensively on these arguments.

With respect to the first of these two points, it needs to be said that an obvious criticism of that argument, which has already been mentioned in section 4.5.3 when presenting the Nash social welfare function, is that the final equilibria and the qualitative implications that can be drawn from it are crucially sensitive to the particular specification of the normative criterion that is adopted. However, there are two further considerations that can now be made. The first is that it could be easily shown that the result obtained is not exclusively associated with the Nash function. For instance, if the relevant criterion were that of inequality aversion, which has received considerable attention in recent works in Behavioural Economics as a relevant motivational prompt to action for individuals, than the structure of the normative criterion would clearly be similar to that associated with the Nash function. This is represented in Table 4.3:

	Co-operation	Defection
Co-operation	*0*	*$(\delta-\beta)$*
Defection	*$(\delta-\beta)$*	*0*

Figure 4.10

Even in this case, the two outcomes associated with Co-operation by one player and Defection by the other would receive a negative evaluation from the normative standpoint, as clearly they would be the outcomes bringing about higher inequality. Moreover, there would be a threshold in the opponent's average play such that the action maximising the normative criterion would switch from Defection to Co-operation as such a threshold would be overtaken. The only difference would lie in that mutual Co-operation and mutual Defection would receive the same evaluation, joint Co-operation not being the only best outcome in terms of the normative criterion. However, this would not prevent the overall situation from being structurally similar to that of the Nash welfare case presented in the previous section, with two pure strategy equilibria associated with mutual Co-operation and mutual Defection,

and a mixed strategy one, which would also turn out to be dynamically stable under the VK criterion.

However, it is also true that different notions for the normative criterion would engender different results. For instance, if a utilitarian conception were used[8], than the normative assessment of the situation would be given by the matrix represented in Figure 4.11.

	Co-operation	Defection
Co-operation	2γ	$\delta+\beta$
Defection	$\delta+\beta$	2α

Figure 4.11

In this situation, under particular configurations of the parameters, it could be the case that the best action in terms of the utilitarian criterion is to Co-operate when the other agent Defects, and *vice versa.* This would bring about what we have called the *exploitative* outcome the one that is preferred in terms of the normative principle. Intuitively, this would occur when the gain in utility attributed to the agent who Defects is higher than the loss in utility suffered by the agent who Co-operates[9]. For the utilitarian criterion attaches the same weight to each agent, and it does not attribute penalties to situations where some agents incur losses with respect to pre-defined allocations, or when inequality arises.

This example shows how the result of ruling out exploitative equilibria does not depend on the general shape of the other-regarding utility, but on the particular

---

[8] Similar result would obtain with a Nash social welfare function where the status quo is set at the origin of the plan.

[9] Not surprisingly, a sufficient condition for this case to arise is $\eta>0$. In fact, this implies that the marginal gain assigned to the defecting agent exceeds the marginal utility forfeited by the co-operating agent. More precisely, the condition that makes Co-operation the action maximising the utilitarian criterion when η is positive is: $p_j < \frac{\delta+\beta-2\alpha}{2\eta}$. Moreover, if the more stringent condition that $\beta+\delta>2\gamma$ is also added, then the above threshold is smaller than one, thus making Defection the action fulfilling the utilitarian criterion when the opponent Co-operates with probability exceeding such a threshold. Hence, the utilitarian criterion may well imply, for some configuration of the parameters, that the exploitative outcome is better from the normative standpoint than mutual Co-operation.

criterion that agents use to make normative evaluations. Since the contractarian criterion preserves some minimal individual 'rights' through the assignment of 'negative' evaluation to outcomes harming agents with respect to their *status quo* allocations, then this type of outcome is prevented from arising as an equilibrium. But under different criteria, these situations can emerge as stable equilibria of the game.

However, the interpretation of the exploitative result now differs from that triggered by normative expectations. It is now the knowledge that self-sacrifice *serves* the public interest that induces people to abide by this norm, rather than merely the fear of others' disapproval. Therefore, one can find a *reason* for her self-sacrificing behaviour that I would call *substantive,* in the sense that it is linked with a *particular* prescription of a normative criterion that is generally endorsed when specific circumstances occur, rather than being justified by a *general* feature of human behaviour that would apply to any situation.

The analogy with the normative expectations theory presented in this Chapter could be made even closer through making not too dramatic changes to the model. In particular, assuming that the two populations of agents have *different* normative criteria, each associated with the material payoffs of the agents of the *other* population, the main implications of that theory would arise within this setting as well. For instance, for an *i*-player, the normative criterion should be shaped as in Figure 4.12.

	Co-operation	Defection
Co-operation	γ	β
Defection	δ	α

Figure 4.12

The compliance with this particular normative criterion would now imply assigning a penalty every time the opponent's payoff is less than what was expected, which is exactly what is prescribed by the resentment hypothesis. Hence, one could see the normative expectations model as a particular case of this framework of analysis, in which each agent sees the other agent's payoff as the normative standpoint of

assessment. The obvious difference lies in that in Sugden' s model there is no *shared* view of morality, as each agent takes the interest of the other as such standpoint[10].

The final aspect that is worth stressing in this final summary is the relevance of a dynamical analysis for the understanding of the nature of social norms. By emphasising the difference between dynamically stable and unstable equilibria, I hope to have shown the limitations of normative expectations as a device to select equilibria, and the necessity to look at other substantive causal elements in order to sustain a social norm. This has been identified with a shared idea of morality that satisfies the requirements of impersonality set out in 1.4. In this way, a social norm can be thought of as being sustained not only by the resentment elicited in the case of its breaching, but also by the idea that agents can perceive that they are endorsing a course of action that is moral as prompted by the adoption of an impartial standpoint.

I hope that with the model developed in this Chapter a step toward a sound account of the two-way relationship between social norms and individual behaviour has been provided. On the one hand, the investigation of the stability of social norms, and of the dynamical process that brings them about, shows how norms are ultimately the result of individual action, and that the lack of motivational force is a cause of their instability. Moreover, the model also allowed for the fact that norms, understood now as common patterns of assessment affecting other-regarding motivations, rather than well-established regularities of behaviour that are the long-run result of the dynamic process of interaction, have an impact on individual behaviour during the very process of their emergence. Norms and individual behaviour are then mutually inter-connected, and the emphasis on dynamical analysis is justified by the necessity to take such reciprocal relationship into account.

---

[10] Unless one wanted to reconsider the exploitative equilibria that we had found in section 4.2.1 as the resulting outcome of the identification of the 'public interest' with the interest of the dominant class. However, this would cast a rather bleak light on this approach

SECOND PART

GROWTH WITH BOUNDEDLY RATIONAL AGENTS, NON-INSTANTANEOUSLY CLEARING MARKETS, AND COMPETING TECHNOLOGIES

INTRODUCTION

In the past few years, the search for new insights in growth theory has intensified under the pressure of the so-called convergence controversy, i.e. the empirical debate over the pattern of convergence, if any, of per capita income levels across countries. After common agreement has been reached over the rejection of the absolute convergence hypothesis (e.g. Barro (1991)), efforts have focussed on testing the validity of the conditional convergence hypothesis, which maintains that per capita incomes of countries sharing the same structural characteristics, e.g. preferences, technologies, population growth and government policies, etc., should converge to the same level regardless of their initial conditions. Evidence has been gathered that conditional convergence holds, but is slow, being about 2% per year (e.g. Mankiw, *et al.* (1992), Barro and Sala-i-Martin (1995)).

However, not only do some researchers question the validity of such evidence, calling for the use of different econometric techniques (e.g. Durlauf et al. (2001), Quah (1997)), but they also doubt that this is the really interesting issue to address (e.g. Azariadis and Drazen (1990), Quah (1996)). When we face such striking differences in per capita incomes across countries as we observe in the real world, the dramatic question is not whether or not poor countries converge to their own level of steady states, but what causes such steady state levels to be so low. Consequently, a sense of dissatisfaction with the neo-classical Solow model, in which steady states are

essentially exogenously determined, led to various refinements of both the neo-classical and endogenous-growth approach. What all of these refinements have in common is the notion of multiple steady states, i.e. the simultaneous presence of two, or more, equilibria, one of which involves a long-run per capita income growth rate that is lower than the other(s). This equilibrium may be called a poverty trap[1], and its empirical counterpart is the so-called club-convergence hypothesis: per capita income of countries that are identical in their structural characteristics converge to the same level provided that their initial conditions belong to the basin of attraction of the same steady-state equilibrium (Galor (1996)).

There are many reasons why multiple steady states may occur in one-sector models of growth. In a Solow-setting, it suffices to introduce heterogeneity across individuals, and different propensities to save out of interest and labour income in order to imply multiple steady states (Galor (1996)). In overlapping generation models, the role of human capital has received much attention. On the one hand, some authors have stressed the social increasing returns to scale from capital accumulation, either because of the positive externalities brought about by individual human capital (Lucas (1988)), or because of some threshold effects in technical progress (Azariadis and Drazen (1990)). Others have focussed on the constraints on individual capital formation stemming from capital market imperfections, especially in the presence of income inequality (e.g. Aghion *et al.* (1999), Aghion and Bolton (1997), Galor and Zeira (1993)), or local externalities (e.g. Durlauf (1996), Benabou (1996)). Closely related is the issue of the impact of financial institutions on growth (e.g. Banerjee and Newman (1998)). Another account hinges more directly on the distribution of income, especially through politico-economic channels: higher inequality lowers the median income position thus bringing about higher tax rates and lower growth (Persson and Tabellini (1994)); or, an initially low level of wealth may generate social conflict, thus hampering the chances of catching up (e.g. Benhabib and Rustichini (1996)). Other

---

[1] There exists another notion of poverty trap in the literature, which is associated with the persistence of an individuals position in the income distribution rather than with aggregate variables. Under this definition, a poverty trap is a state in which dynasties starting out with income below a threshold, converge to a low-level of income, whereas others converge to a high-level (e.g., Moav, 2002, Durlauf (1996)). However, in this paper I will always refer to the concept of poverty trap given in the text.

causes of multiple steady states include endogenous fertility (e.g. Galor and Weil (1996)).

As emphasised by Bernard and Jones (1996), however, very few of the explanations provided focus on technology, despite its undoubted importance for growth, but rather insist on capital accumulation as the privileged factor in accounting for income disparities: in the words of Romer (1993), 'object gaps' seem to count more than 'idea gaps' for students of development. However, some studies have emphasised how technology gaps amongst countries do seem to occur in reality (Bernard and Jones (1996)), and have put forward theoretical explanations for this fact, which hinge on a country's institutional and economic structure as possible barriers to the adoption of the 'leading-edge' technology (Parente and Prescott (1994)). The so-called 'appropriate technology' approach has systematically taken this view, emphasising the necessity of a 'good match' between the technology adopted and the specific structural characteristics of an economy (e.g. Basu and Weil (1998)). For instance, if the frontier technology is produced in advanced countries, and this is designed for the use by skilled workers, then developing countries, which can only rely on unskilled workers, cannot exploit technical advances and bridge the gap with the most advanced ones (Acemoglu and Zilibotti (2001)).

Another strand of research has modelled technology and industrial structure as key factors in orientating the growth path of an economy. In particular, the role of demand spillovers across the various sectors of an economy has been indicated as crucial in making it possible the 'take-off' from a state of low to one of high growth (Murphy, *et al.* (1989)). Analogously, focussing on the supply side, localized technological complementarities amongst industrial sectors can determine multiple steady states, and the presence of some leading sectors can again trigger a take-off process (Durlauf (1993)). This literature has the merit of focussing on the structural aspects of the productive system as possible causes of the macroeconomic performance of the economy and of the steady state selection. But it also has the merit of having prompted a debate as to whether 'history' or 'expectations' are the main determinants of steady states selection – see, e.g. Krugman (1991). The former account states that, because of 'Marshallian' externalities that induce increasing returns

to scale at the sectoral level, the steady state to which an economy converges depends on the 'initial state' of the economic system, e.g. on the degree of sectoral specialisation occurring at the 'beginning' of the 'story'. In other words, within a multiple steady states framework, the economy will follow a (deterministic) convergence path determined by the initial position of the system. This will in fact be the approach that will be pursued with the present model.

The latter account recognises, as well as the former, that there may exist costs and sluggishness in the process of economic agents switching their sector of activity, in particular because of conversion costs of their (human) capital (e.g. Matsuyama (1991)). If, at the limit, such costs are assumed to be infinite, then agents, who are assumed to have *perfect foresight*, will have to take into account the expected discounted sum of future payoffs in either sector in order to decide in which of them to position their activity. But, in the face of sectoral increasing returns to scale and externalities, the expectation on which sector *other* agents locate their activities becomes fundamental for everyone's choice. From this derives the 'expectational indeterminacy' of this approach.

Interesting though it is, this second approach to steady state selection within multi-sector models of growth crucially relies on the assumption of an agent's perfect foresight over her whole time horizon. I believe that this assumption is untenable, at least for the most genuinely interesting economic problems. In fact, given the high non-linearity and complexity of the economic system, the 'beginning of the story', that is, the time in which agents should make their decision as to which sector specialize in, will be characterised by an extremely changeable environment, which could even be likened to a *chaotic* motion. This will indeed be apparent from the diagrams that I report later on in this chapter. Under these conditions, predictions over the future are, almost by definition, impossible, because even slight changes in the initial conditions bring about relatively large changes in the path the economic system will follow. In my view, it is pointless to assume that agents base their decisions over the whole future, and even less to assume that their predictions are, on average, correct. This, at best, may occur in the long run, but by definition the problem we are interested in studying is one of very short run. In fact, only after the 'transition' phase has gone by, and the

economic system has settled on a relatively 'stable' path of convergence towards an equilibrium, can predictions over the future be undertaken with no fear of the 'turbulences' typical of the chaotic motion. The hypothesis of perfect foresight can now be taken as a plausible description of agents' behaviour, but at the cost of neglecting the preceding phase, which is obviously the key one to determine the equilibrium steady state to which the system will converge. This implies that between the two alternatives set out by Krugman, 'history', rather than 'expectations' will be privileged as main explanatory factors.

For the short-term period[2], which is the crucial one to determine the long-run behaviour of the economic system, the use of the alternative 'evolutionary' approach seems more sensible. This makes of *bounded rationality* and *disequilibrium* analysis its two main assumptions (e.g. Nelson and Winter (1982), Hodgson (1999)). Bounded rationality can in turn be broken down into three postulates: (a) myopic expectations; (b) limited information, and (c) limited cognitive abilities. Hence, the postulate of forward-looking behaviour over the whole time horizon is replaced by an assumption of very simple adaptive expectations, which are formed *solely* over the next period, and where the future value of the relevant variable is taken to be the same as the current one. As for assumption (b), in the model I develop, agents are assumed to keep on performing the same action *until* some additional information comes about suggesting that a change in their behaviour is profitable. Consequently, agents will make decisions only on the basis of their *current* payoffs, by comparing the difference between between the action they are presently using and an alternative one, rather than on the expected value of the sum of current *and future* payoff differentials. The way such information becomes available is not formalised in the model, and can be thought of as stemming from a slow process of diffusion of information and of imitation of other agents active in the system. Moreover, given the extreme variability of the system, agents may make mistakes in 'de-codifying' the information they receive, thus implying

---

[2] Notice that the 'start' of 'history' is typically associated with the occurrence of a shock that alters the equilibrium position previously reached. Needless to say, if the environment is so changeable that shocks occur with high frequency, then the system will never actually happen to settle on a stable path, thus making it difficult for agents to form forward-looking expectations. This makes the use of evolutionary modelling even more compelling.

that they could stay put on their current action although informed of an alternative that is 'objectively' better, but subjectively not recognised as such.

As for the non-instantenous market clearking hypothesis, its underpinnings are the same as those supporting the bounded rationality one. In an extremely variable world, commodities are exchanged even *before* the market-clearing level has been reached, thus calling for a disequilibrium type of analysis.

Hence, the two key 'closing' conditions of 'mainstream' economics, i.e. optimal behaviour and market equilibrium, will be replaced by two *dynamical* conditions of *selection of optimal behaviour* and *price adjustment.* The former consists of a rule of motion that 'selects' the best strategies available at the moment a choice is made, the main idea being that in a 'chaotic' world it takes time for an agent to correctly decipher the information and the environment in which they live, so that only a fraction of agents adjust to the currently more profitable strategy (Dosi and Nelson (1993)). Elements of slow learning and imitation of agents of greater success can also be thought of as underpinning of this idea. In particular, a *replicator* dynamics (Weibull (1995)) will describe the aggregate rule of motion of agents' actions. As for the rule of motion of prices, a simple dynamics in which prices rate of change depends on the imbalances between supply and demand in a market will be adopted. The resulting picture is thus one in which the process of adjustment towards market equilibrium and individual optimality is only gradual, and the fact that exchanges take place outside equilibrium causes such equilibrium to vary continuously over time. Agents form *myopic* expectations, and will stick by simple rules of thumb in making decision. Only in the long run, after the aggregate behaviour of the economic system has become sufficiently 'stable', can individuals be thought of as being maximising their payoffs, although the steady state may be sub-optimal. However, due to the non-smoothness of the aggregate 'production function', market may be affected by disequilibrium in either capital or labour even in the long run.

Of course, such hypotheses are in sharp contrast with the standard assumptions of optimising agents and market equilibrium that can be found in most of 'mainstream' economics. A deeper methodological discussion than that sketched out above clearly lies beyond the purpose of this thesis, thus I shall be content with the references

already mentioned. Nevertheless, some tentative comparisons between the two approaches can be advanced even within the present setting. In fact, since the speed of adjustment towards the optimal or the equilibrium level depends on a pair of parameters, the situation of *infinite* speed of adjustment of both prices and individual behaviour, which can be associated with the underlying hypotheses of mainstream economics, can also be analysed as a special case. However, the comparison between the two approaches cannot be said to be complete, since the hypothesis of forward-looking behaviour and perfect foresight that usually accompanies the rationality assumption in most of mainstream economic models will be here replaced by one of simple adaptive expectations.

After the methodological framework has been clarified, I turn on the substantive issues of the analysis. The main goal is to develop a model of growth where technical change and the structural conditions of an economy occupy the centre-stage of analysis as the main dynamic engines for the economic system. Both neo-classical and 'new' Growth Theory generally deals with technical change of the general purpose type, i.e. one capable of affecting the *whole* set of productive techniques. Indeed, it is this assumption that justifies the common representation of technical change as a uniform shift of the isoquants of the production function towards the origin. Instead, I take on the notion of localized technical change, which was originally put forward by Atkinson and Stiglitz (1969) and Salter (1969), and which has also been recently adopted by both the 'appropriate technology' approach and the 'evolutionary economics' approach. The basic idea behind localized technical change is that an innovation is generally capable of affecting only a limited subset of techniques, with general purpose technologies being a limiting case. This corresponds to the shift of some segments, or even single points, of the isoquant towards the origin, rather than taking for granted that the entire isoquant shifts.

Furthermore, I will focus on a structural factor often overlooked in both evolutionary and mainstream approaches - namely the composition of the workforce in terms of its skill attainment. More precisely, labour is not considered *homogenous*, that is directly employable into all of the productive activities; rather, the different technologies are constrained to employ only labour with a particular level of skill. In

particular, two technologies are available, one of which is *skilled* whereas the other is *unskilled* labour intensive. Technical change is then made up of an exogenous and an endogenous component. The former advantages the skilled-intensive technique, whereas the latter makes individual productivity growth rates depend on the degree of concentration of economic activities in that technique. The reason is that technical knowledge is a public good at the individual technique level, but not at the economy-wide level, since knowledge spillovers can occur only amongst firms adopting the same technique.

This is the key characteristic causing the whole process of technical change to be cumulative and determining increasing returns to scale at the sectoral level. Most importantly, this is also what possibly causes the economy to be stuck in a slow-growth trap: although the skilled labour intensive technique *ceteris paribus* brings about faster productivity growth rates, whenever the structural conditions of the economy are initially adverse to its development (e.g. because of skill shortage within the workforce) the economy will concentrate its activities into the other technique, thus precluding the possibility of catching-up. Therefore, in the model, low and high-growth steady states are identified as situations in which all of the productive factors, i.e. capital and labour, are allocated into the unskilled and the skilled intensive technology respectively. On the grounds of local stability analysis and of numerical computer analysis, it is shown that the economic system always converges towards one of these steady states; in particular, the steady state characterised by a balanced growth path between the two sectors can be proven to be unstable.

The main contribution of the paper lies in the analysis of the economic conditions that determine the convergence towards one rather than the other steady state, of the path of convergence, and of the particular characteristics of the steady states thus obtained. In particular, in this setting multiple steady states and lock-in effects arise through channels different from those emphasised in 'mainstream' growth theory, specifically as a failure of the co-ordinating power of the market forces in driving the economic system towards an efficient outcome. The model is general enough to accommodate many facts, such as the striking lack of convergence by poorest

countries with respect to more advanced ones, but also the different experience of the so-called 'convergence clubs' within groups of more homogenous countries.

The resulting model has a flavour similar to evolutionary work dealing with the question of the contest between technologies, mainly that introduced by Arthur (1985), within a urn-scheme stochastic process, and also Durlauf (1993) and Corradi and Ianni (2000) in the context of a spatial-temporal analysis. These models prove the non-ergodicity of stochastic systems, thus presenting multiple equilibria and lock-in of inefficient outcomes as possible long-run steady states. It is also similar to other works that analyse the rise and extinction of different technologies and their impact on the macroeconomic performance of the economy, such as Nelson and Winter (1982), Verspagen (1993), Silverberg and Verspagen (1994). However, the present model is distinct from many of these contributions in that the final outcome is not merely due to random factors, but to underlying causes in labour markets dynamics and the pattern of technical change.

The model also relates to Baumol's analysis of the unbalanced development of the economy (1967), although in his approach the demand side of the economy is seen as the crucial determinant of the unevenness of the growth path, whereas in the present model I focus on the supply side. Furthermore, each of the sectors that composes the economy is modelled like Goodwin's single-sector model of growth (1967). Finally, some links with the 'appropriate technology' approach are also evident, as the slow-growth steady state can be interpreted as a mismatch between the structure of the economy, and in particular its workforce composition, and the advanced technology.

The basic structure of the model is introduced in section 5.1: this is composed of two parts, one related to the sphere of production, where the notions of the pair of available technologies (section 5.1.1), of their productivity growth rates (section 5.1.2), and of the rules of capital accumulation (section 5.1.3) are outlined. The second part describes the dynamics of the labour market (section 5.1.4). Given the off-equilibrium nature of the model, it is also necessary to specify the behaviour of the economy in the recurrent situations of imbalances and rationing in the market (section 5.1.5).

The first version of the model, characterised by the impossibility of workers changing their initial type of skill, is presented in section 5.2. After an illustration of

the economic forces driving the evolution of the model (section 5.2.1), an analysis of the local stability of its possible steady states is carried out (section 5.2.2). This turns out to be inconclusive, given the non-generic nature of the system of differential equations. Therefore, the results of a numerical analysis of the evolution of the system is presented. Not only does this allow us to discuss the final outcome of the evolution of the model, in terms of convergence towards one of the two technologies, but also it enables an analysis of the path actually followed by the economy along this process. Section 5.3.1 presents a scenario of lock-in towards the unskilled intensive technique when the skill shortage in the workforce is particularly marked, while section 5.3.2 shows how this result is reversed when the initial skill composition is even slightly modified in favour of skilled labour.

Chapter 6 studies a second and more general version of the model that allows the possibility of mobility of the workforce between the two sectors. In section 6.1 the presence of mobility costs for the workers is modelled, and section 6.2 shows how both outcomes of convergence towards the techniques are possible. In particular, by means of numerical analysis, in section 6.3 I show how reducing the mobility costs associated with skill upgrading it is possible to reverse the result of convergence to the low-growth steady state into that of convergence to the high-growth one. Section 6.4 draws out some conclusions, leaving a discussion of the mathematical details of the model to section 6.5.

CHAPTER 5

GROWTH WITH LABOUR RIGIDITY

5.1 THE SETTING OF THE MODEL

5.1.1 Production

Consider a market for a homogenous good Q, where *two* different *techniques*, differing upon their labour skill-intensity, are available to firms. For simplicity, I suppose that the first technique exclusively adopts *skilled* labour, while the second only employs *unskilled* labour. Each technique is uniquely determined by the pair of coefficients expressing the requirements of capital and labour per unit of output, so that a different pair of coefficients of production *per se* implies a different technique[3]. Nevertheless, as I shall elaborate in the next section, although the production coefficients of a technique are fixed in an instant of time, they may change over time as an effect of technical change. The two techniques of production can thus be represented by a fixed coefficient Leontief production function:

$$Q_i = \min\left\{a_i L_i, \frac{K_i}{c}\right\} \quad i = 1,2 \tag{5.1}$$

where L_1 and K_1 represents the employment of skilled labour and capital in the skill-intensive technique, and L_2 and K_2 the amount of unskilled labour and capital employed in the unskilled-intensive one. c is the fixed coefficient of the content of capital in one unit of output, assumed to be equal for the two technologies, and a_i is labour productivity. Obviously, total production is given by: $Q=Q_1+Q_2$.

This model can also be thought of as representing the whole of the economy. In this framework, Q is the aggregate bundle made up of two commodities produced in

---

[3] Dealing with only two techniques of production, rather than the continuum of techniques that would make up a typical 'neoclassical' isoquant of production, precludes marginal substitutability of inputs, enabling firms to operate only a discrete choice between the two available techniques. This characterisation is consistent with indivisibilities in the production process (Antonelli, 1995).

the two main industrial *sectors* of the economy, which employ techniques, or, rather, *technologies*[4], differing as to their labour skill-intensity. The sector employing the skilled labour intensive technology thereby represents an advanced hi-tech industry, whereas the other is a 'traditional' low-tech sector. In order for Q to be a sound concept in this setting, one also needs to assume that the relative prices of the goods remain constant over time, for instance because of infinite elasticity of their demands. In a neo-classical framework, free entry into each sector and arbitrage conditions guarantee that the sectors grow at equal rates, thus making it possible to abstract away from individual outputs and to consider the aggregate of the two. In the present context, however, given the off-equilibrium approach, it may be possible that sectors grow at different rates even in a steady state, a characteristic which is usually referred to as unbalanced growth.

Given the close correspondence between the two cases presented, in the following sections the terms technique and technology will be used interchangeably, and the term *sector* will represent at the same time the segment of a market employing one of the two techniques in the single market version of the model, or the industry adopting a certain technology in the aggregate version.

5.1.2 Technical change

As pointed out in the introduction to the second part, the account of technical change that will be developed is of the localized type; that is, two independent laws of motion of technical advance will be set out for the two different techniques. Furthermore, technical change will depend upon an endogenous and an exogenous factor. Let us deal with the first factor.

Some notable economists, (e.g. Arrow (1962), Paul Romer (1994)), have emphasised the non-rival nature and the -at least partial- non-excludability of innovations. In fact, although there are many ways whereby firms can temporally appropriate the benefits deriving from an innovation, we can expect that in the long run a large amount, if not all, of the knowledge associated with a technical innovation

---

[4] The term technology seems more suitable to dealing with industrial sectors rather than technique. However, I will consider the two as substantial synonymous.

will be spread over the rest of the economy. An opposite view is taken by economists of the evolutionary area (e.g. Dosi (1988)) who argue that the nature of technological knowledge is largely tacit, thus, at least to some extent, appropriable by firms. According to this view, even in the long run can first-innovators still keep their 'technological lead' from the followers. Consequently, the scope of the phenomenon of technical spillovers throughout the whole of the economy would be limited.

In the present specification I take on an intermediate position between those two views, arguing that technical information spillovers take place at the level of the individual technique, but not at the economy-wide level. In other words, technical information is a public good only at the technique-specific level, in particular because innovations carried out by a firm adopting a certain technique can be imitated by firms employing the same technique, but not from those using the other. For simplicity, the imitation process is assumed to be instantaneous. An analogous interpretation of this idea would be that firms belonging to the same sector of the economy create a 'net' through which technical information is transmitted amongst them (e.g. Antonelli (1995), Maurseth and Verspagen (1999)).

No explanation is given in the model about how and why such innovations are carried out, and the R&D variable is then omitted. However, it is assumed that the larger the share of firms active in a certain sector, the higher the probability that an innovation is carried out. That is, applying the large numbers law, the growth rate of technical change depends positively on the density of activity in a certain sector, which is measured by the share of capital invested therein. Finally, technical change is assumed to be of the labour-augmenting type.

The following equation summarises all these considerations. It describes the rule of motion of labour productivity for each technique:

$$\frac{\dot{a}_i}{a_i} = g_i \kappa_i \tag{5.2}$$

where κ_i is the share of capital invested in technique i:

$$\kappa_i = \frac{K_i}{K} \tag{5.3}$$

g_1 and g_2 are fixed parameters satisfying the following condition:

$$g_1 > g_2 \tag{5.4}$$

The fact that a_i depends on κ_i captures the *endogenous* factor of technical change: the larger the concentration of economic activity in technique i, the larger the technique-specific innovation rate[5,6]. The parameters g_i, instead, portray the *exogenous* factor of technical change, which is usually thought of as being linked with the rate of advance in scientific discoveries. In fact, many scholars of the subject emphasise the essential role of growth in general scientific knowledge to bring about technological innovations: this acts as the ultimate determinant and as a constraint to the technological knowledge present in an economic system (Rosenberg (1982)). The fact that science mainly evolves independently from economic activity makes it an exogenous factor of development[7]. Condition (5.4) takes account of the evidence that skilled labour intensive techniques have in the last decades been more efficient than unskilled intensive ones, particularly because of the complementarity between information technology and skilled labour (Berman *et al.* (1998)). In the present model, this is equivalent to saying that scientific discoveries are more easily applicable to technique 1. This implies that technique 1 has a higher *potential* for growth, in the sense that economic growth is higher when productive factors are entirely allocated in the skilled intensive technique rather than in the unskilled intensive one. Therefore, steady

---

[5] In some preliminary empirical analysis, I have collected some evidence confirming this relation: in particular, the manufacturing sectors have been classified in four groups on the grounds of their technological intensity, in accordance with OECD (1997). Hence, for a sample of OECD countries, the productivity within each group, normalised for a country's average productivity, can be shown to depend positively on their value added share, with respect to manufacturing value added. This supports the view that economic systems *specialising* in a particular technology, or industrial sector, in terms of their value added share, also experience higher productivity growth rates.

[6] It can be noticed that the specification of the endogenous component of technical progress comprises a *negative* externality between firms active in different sectors: not only does a firm leaving a technique for the other increase the productivity in the new sector, but also it decreases that of the previous sector. Indeed, a relation consistent with the theoretical considerations previously set out should make the productivity growth dependent on the *absolute* level of capital present in a sector, rather than on its *relative* value. However, the specification adopted can be justified on grounds of analytical tractability, and in any case it does not affect the results of the model, since, as we shall see in section 5.1.3, the choice of the firm as to their location in the technology scale is based on the *relative* profitability of the two techniques.

[7] I proved that other specifications of the productivity equations, e.g. a linear equation in g and κ, lead to the same results as the one adopted; in general, though, I cannot provide a general proof of the robustness of the results.

states characterised by entire allocation of factors within the skilled-intensive technique will carry out higher growth rates than alternative steady states.

5.1.3 Capital accumulation

5.1.3.A Rule of Motion of Aggregate Capital

Since two productive sectors are present, we need two different laws of motion of capital: one refers to the aggregate level of capital accumulation whereas the other shows how capital distributes between the two sectors. In what follows, we shall assume that there exists a *continuum* of firms, so that the *dimension* of each of them is negligible. Moreover, each firm possesses one (infinitesimal) unit of capital, which can be invested in either technique. The following are aggregate relations, which can nevertheless be accounted for in terms of sensible individual behaviours. Let us first introduce some variables:

$$K=K_1+K_2 \tag{5.5}$$

$$r_i \equiv \frac{\pi_i}{K} \equiv \frac{Q_i - w_i L_i}{K_i} = \frac{1-\frac{w_i}{a_i}}{c} \tag{5.6}$$

$$\bar{r} = \kappa_1 r_1 + \kappa_2 r_2 \tag{5.7}$$

Equation (5.5) defines the aggregate level of capital as the sum of the amount of capital invested in each technique. r_i describes the individual profit rate, and $\bar{r}$ is the overall rate of profit obtained by weighing the individual profit rates with the respective capital shares. Let us call y_i the cost of labour per unit of output produced:

$$y_i \equiv \frac{w_i}{a_i} \tag{5.8}$$

Then, we can write a compact expression for the technique-specific rate of profit:

$$r_i = \frac{1-y_i}{c} \tag{5.9}$$

The share of total labour income is thus given by the following expression:

$$\bar{y} \equiv \kappa_1 y_1 + \kappa_2 y_2 \tag{5.10}$$

As for consumption and investment behaviour, in order to avoid technical complexities I simply assume that at each instant of time workers spend a constant share of their income, for simplicity set equal to one, in consumption, and firms reinvest a constant proportion of their profits, assumed equal to one as well[8]. Therefore, the overall flow of investments coincide with the amount of profits in a period, and the proportional growth rate of aggregate capital is equal to the rate of profit:

$$\frac{\dot{K}}{K} = \bar{r} = \frac{1 - \kappa_1 y_1 - \kappa_2 y_2}{c} \tag{5.11}$$

5.1.3.B Rule of Motion of Sectoral Capital

I now specify the macroeconomic evolution of the system for what concerns the distribution of capital across technologies, based on the assumptions of boundedly rational behaviour at the individual level. The specification that will be adopted draws on a model that has been put forward in the evolutionary literature (Silverberg and Verspagen (1994)), which captures the idea that adjustments towards optimality can occur only slowly because of informational and/or cognitive constraints on the agents. Equation (5.12) describes the sectoral growth rate of capital:

$$\frac{\dot{K}_i}{K_i} = r_i + \alpha(r_i - \bar{r}) \qquad \alpha > 0 \tag{5.12}$$

This equation is similar to equation (5.11) in its first component, in that capital accumulation in sector i depends on the amount of profits made by firms active therein, which, in accordance with the behaviour assumed in the previous section, are immediately reinvested: this is the 'normal' rate of accumulation r_i, that is. However, the second component takes into account the possibility that firms switch towards the technique that is currently more profitable, thus making its growth rate higher. Adjustment costs are assumed to be negligible. Such a second component is consistent

---

[8] Notice that nothing would be lost by setting the propensities to consume and invest at a level less than one. This assumption is common to models of the Kaldorian tradition, which, more generally, assumed that the saving propensity of firms was higher than that of consumers. In models of neo-classical growth, the same result of constant propensity to save obtains, but in that case the horizon span is infinite.

with a bounded rational behaviour, in that only the extreme case of α equal to infinity does correspond with the situation of instantaneous adjustment to optimality. In all of the other cases, at each instant of time only a part of the firms switch to the more advantageous technique, at a rate that is proportional to the extent of the profitability difference. Therefore, α may be seen as an index of the 'amount' of information available to agents. Instead, the term expressing the relative profitability between techniques is justified on the grounds of agents' cognitive limits: a higher difference in relative profitability implies a faster flow of firms, the idea being that firms can in this case spell out more easily the available information and thus move in the 'right' direction[9].

Under this general framework, various microeconomic accounts of firms' actual individual behaviour and of the extent of cognitive and informational constraints are possible[10]. An interesting characteristic of the rule of motion set out in equation (5.12) is revealed by expressing it in terms of the share of capital:

$$\frac{\dot{\kappa}_i}{\kappa_i} = \frac{(1+\alpha)(1-\kappa_i)(y_j - y_i)}{c} \tag{5.13}$$

It is now evident that a version of the *replicator dynamics* is implicit in the rule of motion of firms across techniques (Weibull (1995)). In fact, (y_j-y_i) is nothing but the difference in the rates of profit[11]. We can thereby conclude that the dynamics of the motion of

---

[9] To be sure, the two aspects of information and cognitive ability are interrelated between one another: agents with more sophisticated cognitive abilities will also have more incentive in collecting information, thus influencing the magnitude of the parameter α. In any case, a value of α equal to infinity will be taken to represent the limit situation of fully informed and perfectly optimising agents who immediately recognise in each period which is the best technique, and move towards it.

[10] A second account that would be consistent with equation (12) would be one in which fully informed and perfectly rational agents, though unable to make forecasts over the future so that their horizon only spans the current period, face delays in the transfer of capital from one sector to the other. In this case, α would then measure the speed at which capital can be scrapped in one sector and reinvested in the other sector.

[11] Not surprisingly, the argument can be easily restated in the language of Game Theory: α, the parameter relative to the degree of information present in the economy, may be seen as depending on the probability of being chosen at random from the group of firms and matched with another firm. The difference (y_j-y_i) is the differences in the payoffs currently earned by the two 'players', thus representing the incentive to adopt the alternative technique in the circumstance it is performing better: in this case the alternative strategy best fits the environment, meaning a higher rate of 'reproduction' of the technology.

firms across sector is grounded on the idea that firms *imitate* the agent of greater success on the basis of a slow process of diffusion of the information and limited cognitive abilities.

5.1.4 Labour Market

Let L^S define the overall level of the workforce present in the economy. Assuming that the population does not grow over time, we can with no loss of generality normalise this level to 1. The population is made up of two types of workers, skilled and unskilled, whose level is indicated by the variable s and $(1-s)$ respectively. In this section I assume that workers cannot move across sectors, and this, together with the assumption that every worker supplies all of her endowment of labour, makes the level s fixed. It is helpful to introduce some notation for labour supply in a specific market:

$$L_i^S = \begin{cases} s & \text{if } i=1 \\ 1-s & \text{if } i=2 \end{cases} \tag{5.14}$$

Due to the off-equilibrium nature of the approach, a pair of equations describing the motion of prices in response to an imbalance in the market is needed. Labour demand will be determined by the firms' level of capital, which yields the following linear relation:

$$x_i \equiv L_i^D = \frac{K_i}{a_i c} \tag{5.15}$$

In the first instance, we assume a simple linear relation between the proportional growth rate of wages and the imbalances in the labour market:

$$\frac{\dot{w}_i}{w_i} = \gamma\left(x_i - L_i^S\right) \tag{5.16}$$

γ is a parameter expressing the speed with which imbalances on the labour market affect wages: in the hypothesis of instantaneous market clearing, this parameter would be equal to infinity.

However, I will add to this relation another term measuring the impact of redistribution policies that are carried out 'externally' to the market:

$$\frac{\dot{w}_i}{w_i} = \gamma\left(x_i - L_i^S\right) + b_i \tag{5.16'}$$

The parameters b_i may be thought of as the result of a bargaining between the relevant groups of agents present in the economy over income distribution. It is indeed clear that its introduction alters both the distribution of income between labour and profits and between the two wages with respect to that obtained through the market, allowing a sort of 'guaranteed' increase in wages[12]. Therefore, it is perhaps more convenient to express it in terms of the increase in labour productivity, as it happens in 'real' bargaining over income distribution:

$$b_i = \eta g_i \qquad \eta \geq 0 \tag{5.17}$$

Therefore, in this specification parameter b_i represents the share of productivity gains accrued to labour income independently from market interactions. As we shall see, the value assumed by parameter η will be crucial in order to determine whether the evolution of the system reaches a situation of structural unemployment or not.

A further constraint regards profits: the following condition allows for the fact that firms would temporarily shut down their activities when experiencing negative profits:

$$0 \leq y_i \leq 1 \tag{5.18}$$

Clearly, when y_i hits the boundary level of 1, wage growth could not exceed productivity growth, since claims from workers to get wage increases above that level could not be accommodated by firms just breaking the even. This constraint, along with equations (5.2) and (5.8), yields the final expression for the law of motion of y_i:

$$\frac{\dot{y}_i}{y_i} = \begin{cases} \gamma\left(x_i - L_i^S\right) + (\eta - 1)\kappa_i g_i & \text{if } y_i < 1 \\ \min\left\{0, \gamma\left(x_i - L_i^S\right) + (\eta - 1)\kappa_i g_i\right\} & \text{if } y_i = 1 \end{cases} \tag{5.19}$$

The actual level of employment will obviously be the minimum between labour demand and supply. Indicating such variable with L_i, we have:

$$L_i = \min\left\{L_i^D, L_i^S\right\} \tag{5.20}$$

---

[12] The reader may have noticed that in such a specification the wages law of motion is formally equivalent with a Phillips curve relationship, where the NAIRU is given by the expression $L_i^S - \frac{b_i}{\gamma}$.

5.1.5 The behaviour of the system when labour demand is rationed

The explicit introduction of labour market requires an amendment of what outlined above regarding the rules of capital accumulation. In fact, the possibility of having an excess of labour demand over supply implies that a fraction of capital is actually left idle, and of course this will affect both the level of profits and of investments.

More precisely, suppose labour demand is in excess, so that only the fraction of capital that can be matched with labour supply is employed in production. The rest is unproductive, because of the perfect complementarity between labour and capital associated with the Leontief technology. Let K^E_i be the amount of capital *effectively* being employed, while K_i is the overall amount of capital *present* in a sector, but not necessarily employed. K^E_i will be given by the following expression:

$$K_i^E = \begin{cases} ca_i L_i^S & when \quad L_i^D > L_i^S \\ K_i & when \quad L_i^D \leq L_i^S \end{cases} \tag{5.21}$$

More generally, we can define the ratio of capital actually employed over the total as:

$$u_i = \begin{cases} 1 & when \quad L_i^S \geq L_i^D \\ \dfrac{ca_i L_i^S}{K_i} & when \quad L_i^S > L_i^D \end{cases} \tag{5.22}$$

The variable u_i will be called the *capacity utilisation of capital* in sector *i*. When labour supply is in excess there will be no rationing, leading to *full* utilisation of capital, represented by a value of 1 for such a variable. When labour demand is in excess u_i will take on values less than 1. Therefore, in situations of rationing, a percentage (1- u_i) of firms *present* in sector *i* is unable to undertake any productive activity. These firms will offer higher wages than the current one, thus hoping to attract workers currently working for other firms in order to enter the market. This has the effect of raising the level of wages in accordance with equation (5.20).

The eventuality of rationing affects both the levels of profits and investments, as these quantities are determined by the capital that is actually being employed. Therefore, the foregoing expressions for the growth of capital must be corrected to take into account its possible rationing. One finds that the overall rate of profit, which is equal to the overall rate of growth of capital, is now given by:

$$\frac{\dot{K}}{K} = \bar{r} = \kappa_1 u_1 r_1 + (1-\kappa_1) u_2 r_2 \tag{5.23}$$

where, as stressed above, the rates of profit in each sector are computed in terms of effective capital:

$$r_i = \frac{Q_i - w_i L_i}{K_i^E} \tag{5.24}$$

This is the quantity that firms actually use when comparing their current level of profit and that made possible by the alternative technique. Analogously, the growth rate of capital in each sector will be given by:

$$\frac{\dot{K}_i}{K_i} = u_i r_i + \alpha\left(u_i r_i - \bar{r}\right) = \frac{u_i[1+\alpha(1-\kappa_i)](1-y_i) - \alpha u_j (1-\kappa_{i1})(1-y_j)}{c} \tag{5.12'}$$

which of course boils down to equation (5.12) when both u_1 and u_2 are equal to 1. Equation (5.13) is subject to an analogous change:

$$\frac{\dot{\kappa}_i}{\kappa} = \frac{(1+\alpha)(1-\kappa_i)\left[u_i(1-y_i) - u_j(1-y_j)\right]}{c} \tag{5.13'}$$

5.2 The Steady States of The Model

5.2.1 An economic insight into the model

The dynamics of the model set out above can be represented by a non-autonomous system of differential equations in the six variables representing the dynamics of each sector, i.e. productivity, capital, unit cost of labour. However, letting x_1 and x_2 substitute for K_1 and K_2, and introducing κ_1, whose value is indeed determined by x_1 and x_2, we can keep out productivity and reduce the system to an autonomous 5-dimension system. This is what obtains in the interior of the space:

$$\dot{\kappa}_1 = \frac{(1+\alpha)(1-\kappa_1)[u_1(1-y_1)-u_2(1-y_2)]}{c}\kappa_1 \tag{5.25}$$

$$\dot{y}_1 = [\gamma(x_1 - s) + (\eta - 1)\kappa_1 g_1]y_1 \tag{5.26}$$

$$\dot{y}_2 = [\gamma(x_2 - (1-s)) + (\eta - 1)(1-\kappa_1)g_2]y_2 \tag{5.27}$$

$$\dot{x}_1 = \left(\frac{u_1[1+\alpha(1-\kappa_1)](1-y_1) - \alpha u_2(1-\kappa_1)(1-y_2) - c\kappa_1 g_1}{c}\right)x_1 \tag{5.28}$$

$$\dot{x}_2 = \left(\frac{u_2[1+\alpha\kappa_1](1-y_2) - \alpha u_1\kappa_1(1-y_1) - c(1-\kappa_1)g_2}{c}\right)x_2 \tag{5.29}$$

The dynamics forces driving the model can perhaps be better appreciated if they are considered each in turn:

A) *Productivity*: Productivity levels are one of the determinants of the relative profitability of the two techniques. It is affected by the concentration of firms in a sector because of the positive externalities gained through knowledge spillovers. This process determines a peculiar phenomenon of cumulativeness in technical change, since once an economy 'specialises' in a technique, that is to say allocates a large share of capital in a sector, it becomes increasingly difficult for the other sector to bridge the productivity gap.

B) *Wages*: The intensive use of a certain technique brings about an increase in the associated level of employment, which in turn increases wages and reduce profitability.

C) *Skill shortage*: This factor refers to a *structural* characteristic of the economy, given by the condition of relative abundance of the workforce in each market. As outlined above, if the labour supply associated with a particular technique is relatively scarce then the excess of labour demand will bid wages up.

These three factors can have a counterbalancing effect on each other; particularly, the wage effect and the possible presence of skill shortages of skilled labour might slow down, or even impede the economic system from converging towards the efficient technology.

In more detail, one can notice that the model is symmetric in the pairs of variables (x_1,y_1) and (x_2,y_2). Moreover, when capital is completely allocated in one sector (that is,

$\kappa_1=0$ or $\kappa_1=1$) then the pair of equations that are now relevant 'loses' every link with the other two equations. For instance, when $\kappa_1=1$, then equations (5.26) and (5.28) are 'autonomous' from the other two variables and make up a 'sub-system' of equations that is known as Lotka-Volterra, or predator-prey, model. This two-dimension system of equations has been extensively studied both in the mathematical literature (Hirsch and Smale, 1974) and in Economics (Goodwin, 1967). Its basic characteristic is to display a persisting cyclical behaviour in the two relevant variables (capital and cost of labour), because an excess of labour demand drives wages up, thus reducing the rate of profit and investment. In turn this decreases the level of production and employment, so that wages diminish and trigger a new phase of increase in investments. The virtue of this simple model is that it generates *endogenous* cyclical fluctuations around a trend within a model of growth. Hence, the system under exam looks like a generalisation of the predator-prey model, being a 'combination' of two such models, and boiling down to one of them when converging to the boundary of the κ_1 axis.

5.2.2 Analysis of local stability

The steady states of the system can be divided into three categories: convergence towards the high-growth equilibrium, convergence towards the slow-growth equilibrium, and, finally, a balanced growth path between the two sectors of the economy. For *convergence* to a sector I mean the process that leads asymptotically to the state of complete allocation of capital – and, from the next Chapter, of labour as well – within that sector. So, we shall observe one sector becoming the *leading* one of the economy, as its share in overall production continuously rises, and the other being confined to a *residual* role. The balanced growth path solution, instead, depicts a situation in which the two sectors grow at the same rate.

An assessment of local stability is not possible for the first two categories of steady states, because the presence of some purely imaginary eigenvalues makes the system locally non-hyperbolic (Guckhenheimer and Holmes (1990)). In section 6.5 the through analysis of local stability is reported. Still, the numerical analysis that I have conducted, part of which is reported in section 5.3, shows that all of such solutions

look like attractors of the system under some values of the parameters. Instead, the solution associated with the balanced growth path can be immediately proved to be unstable by local stability analysis.

5.2.2.A High-growth steady states

$$\text{A1)}\ \left\{\kappa_1 = 1 \quad y_1 = 1 - cg_1 \quad x_1 = s - \frac{(\eta - 1)}{\gamma} g_1 \quad y_2 = \frac{1 + \alpha(1 - cg_1)}{1 + \alpha} \quad x_2 = 1 - s\right\}$$

This solution is characterised by complete allocation of capital into the efficient technique. It holds under the condition that η be greater than 1, thus it implies a positive level of unemployment for skilled labour and full employment for unskilled labour. One can also notice that a greater speed of adjustment in labour market, as measured by coefficient γ, helps to reduce the level of unemployment, which at the limit is then equal to zero. Moreover, when α also goes to infinity, which corresponds to the case of perfect information and rationality of the agents (see section 5.1.3.B), the profit rate in sector 1, $1 - y_1$, equates that of sector 2, $1 - y_2$, thus making firms indifferent between the two sectors. This state seems indeed to reflect a typical 'neo-classical' equilibrium, where labour markets clear and all sectors of the economy have the same level of profitability, though all the activities are concentrated in the first sector. Hence, the introduction of non-instantaneous market clearing and limited information within the model brings about structural unemployment and a persistent gap in the two sectors profit rates.

Another steady state characterised by convergence towards the first sector obtains:

(A2)

$$\left\{\kappa_1 = 1,\ y_1 = 1 - cg_1 - c\left(\frac{1 - \eta}{\gamma s}\right) g_1^2,\ x_1 = s + \frac{(1 - \eta)}{\gamma} g_1,\ y_2 = \frac{1 + \alpha(1 - cg_1)}{1 + \alpha},\ x_2 = 1 - s\right\}$$

This was found under the limitation that $\eta<1$, thus it implies full employment of labour and rationing of capital in the first sector, and full employment of both capital and labour in the second technique.

5.2.2.B Low-growth steady states

We also find a couple of steady states symmetrical to those just found, though they are characterised by convergence towards the second sector:

$$\text{(B1)} \quad \left\{ \kappa_1 = 0 \quad y_2 = 1 - cg_2 \quad x_2 = (1-s) - \frac{(\eta-1)}{\gamma} g_2 \quad y_1 = \frac{1+\alpha(1-cg_2)}{1+\alpha} \quad x_1 = s \right\}$$

(B2)

$$\left\{ \kappa_1 = 0, \; y_2 = 1 - cg_2 - c\left(\frac{1-\eta}{\gamma(1-s)} \right) g_2^2, \; x_2 = (1-s) + \frac{(1-\eta)}{\gamma} g_2, \; y_1 = \frac{1+\alpha(1-cg_2)}{1+\alpha}, \; x_1 = s \right\}$$

Solution (B1) is an equilibrium with 'structural unemployment' in the leading sector of the economy, i.e. sector 2, and full employment of both inputs in the residual one; again this solution holds under the restriction that η is greater than 1. Conversely, solution (B2) predicts an outcome with full employment of labour in both sectors, under the condition that η is less than 1. The properties of stability of these steady states are the same as those found out for the case of convergence towards the first sector.

5.2.2.C Balanced growth path

This is the only steady state in which both technologies coexist:

$$\text{(C1)} \quad \left\{ \begin{array}{ll} \kappa_1 = \dfrac{g_2}{g_1 + g_2} & y_1 = \dfrac{g_2 + g_1(1-cg_2)}{g_1 + g_2} \quad x_1 = s - \dfrac{(\eta-1)}{\gamma} g_1 g_2 \\ y_2 = \dfrac{g_2 + g_1(1-cg_2)}{g_1 + g_2} & x_2 = 1 - s - \dfrac{(\eta-1)}{\gamma} g_1 g_2 \end{array} \right\}$$

Notice that such a steady state is not constrained by any limitation to parameter η: it can thus depict both a situation of full employment of labour or of structural unemployment, boiling down to no unemployment when $\gamma=0$. It is easy to show that for this value of κ_1 the productivity remains constant across the two sectors; the other values ensure that labour markets clear, so that there is no tendency for the system to move away from such a configuration. Since both sectors evolve according to the same growth rate, the economy can be said to follow a balanced growth path. Nevertheless, an analysis of its local properties of stability shows that such an outcome

is in fact unstable (see section). The economic reason is to be found in the property of cumulativeness of technical change: if this state is perturbed even slightly, then the sectoral productivity will differ, thus attracting some firms to move to the more profitable technique. As a consequence, the sector that by chance happens to be the more profitable will experience positive sectoral economies of scale that will suffice to break the balance between the two profit rates.

5.3 Analysis of Global Stability

Due to the complexity of the model, a thorough global investigation of the properties of the system is not possible. Therefore, using a specifically designed programme for Maple V, I have worked out a series of numerical analyses to test the behaviour of the system. The interested reader can find some notes in the last section of the Appendix. Overall, the main conclusion one can draw is that all of the above dubious cases of steady states turn out to be stable equilibria of the model for some values of the parameters. In what follows I shall highlight some of their features.

5.3.1 First Scenario: Convergence Towards the Inefficient Technology with High Skill Shortage

First, I consider a situation where the condition of skill shortage is quite marked, as the economy starts off from a position where two thirds of the population are unskilled; that is to say, $s=1/3$. Besides, I also assume that the initial situation is one of initial perfect symmetry for all of the other variables, such that the two techniques are equitably profitable, and firms are equally distributed across them[13]. This should be appropriate for a situation of absolute ignorance over the properties of the techniques at the beginning of the 'story'. Moreover, parameter η, the key to income distribution,

---

[13] The particular values for the parameters have been chosen consistently with works by Silverberg and Verspagen (1994) – for what concerns α, the degree of individual rationality and information - and Barro and Sala-i-Martin (1996) – for the capital output ratio c and sectoral growth rates g_1 and g_2: $\alpha=1$; $g_1=0{,}04$, $g_2=0{,}02$; $c=3$; $\gamma=0.5$; $s=1/3$; $\eta=1$.
The set of initial conditions is meant to depict a situation of even distribution of firms across the two sectors:
$\kappa_1(0)=0.5$; $y_1(0)=0.5$; $y_2(0)=0.5$; $x_1(0)=0.5$; $x_2(0)=0.5$; $a_1(0)=1$; $a_2(0)=1$

is assumed to be greater than one. Therefore, the feasible steady state between those listed in section 5.2.2 would be solution (B1).

The main result is that the system converges asymptotically to the slow-growth steady state, as all capital becomes invested into technique 2 through a series of periodical oscillations that progressively dampen down (Figure 5.1). The reason can be investigated by looking at the behaviour of productivity growth rates. Technique 1 starts off with a higher productivity, as a result of the even distribution of firms across technologies at instant 0 (figure 5.2). However, firms become soon attracted by the possibility of hiring cheap labour in the unskilled labour market, thus boosting technique 2's productivity. Hence, after few periods, the productivity in the second sector leapfrogs the other, and this is sufficient in order to determine a form of technological lock-in towards the second technique.

It is worth noticing that the first sector does not completely disappear over time: as Figure 5.3 shows, production growth rate settles on a 0-growth path on average, in which the flow of firms moving to the more profitable sector is exactly compensated by the flow of new investments into this sector. This is the sense in which we can call this sector the *residual* one of the economy. The overall growth rate tends to stabilise over the same growth rate as that of the leading sector – sector 2, that is - but the individual growth of the individual sectors show much widest fluctuations, which partly offset each other because of their different periodicity.

The subsequent diagrams depict the dynamics within the labour market in the leading sector; that of the residual sector is similar thus it will not be commented. Employment and unit cost of labour settle on a cyclical path (Fig. 5.4). In the long-run, their trajectory converges towards a limit cycle, typical of Lotka-Volterra systems (Fig. 5.5). Notice that the coordinates of the centre of the cycle correspond with solution (B1), once the parameters have been assigned their numerical values[14]. While the first sector periodically experiences phases of full employment, a state of structural unemployment in the leading sector of the economy occurs.

---

[14] Plugging in the value of the parameters, one then obtains that the trajectories should either orbit or converge towards the following set of values: $\{\kappa_1 = 0 \quad y_2 = 0.94 \quad x_2 = 0.64667 \quad y_1 = 0.97 \quad x_1 = 1/3\}$

One may question the reason why the two markets do not clear, even in the presence of flexible prices. In general terms, this is the result of the non-instantaneous adjustment in the labour market (see discussion in section 5.2.2). However, there is a further reason, related to income distribution. Provided that techniques have by construction fixed coefficients of production, firm labour demand will depend on the level of wages only indirectly, by means of the following dynamical mechanism: a low level of wages triggers high profits and then high investments, thus increasing the stock of capital and the demand for labour, and vice versa. However, if the 'institutional' redistributive component is too high, which is the case when η is greater than 1, then capital accumulation will be too low to match labour supply. On the other hand, as we shall observe in the following section, for values of η less than 1 we obtain full employment of labour, but this time it is a portion of capital that is left idle. Therefore, too large a share of income devoted to labour implies too slow capital accumulation, thus creating unemployment. Finally, Figure 5.6 shows the evolution of income shares over time.

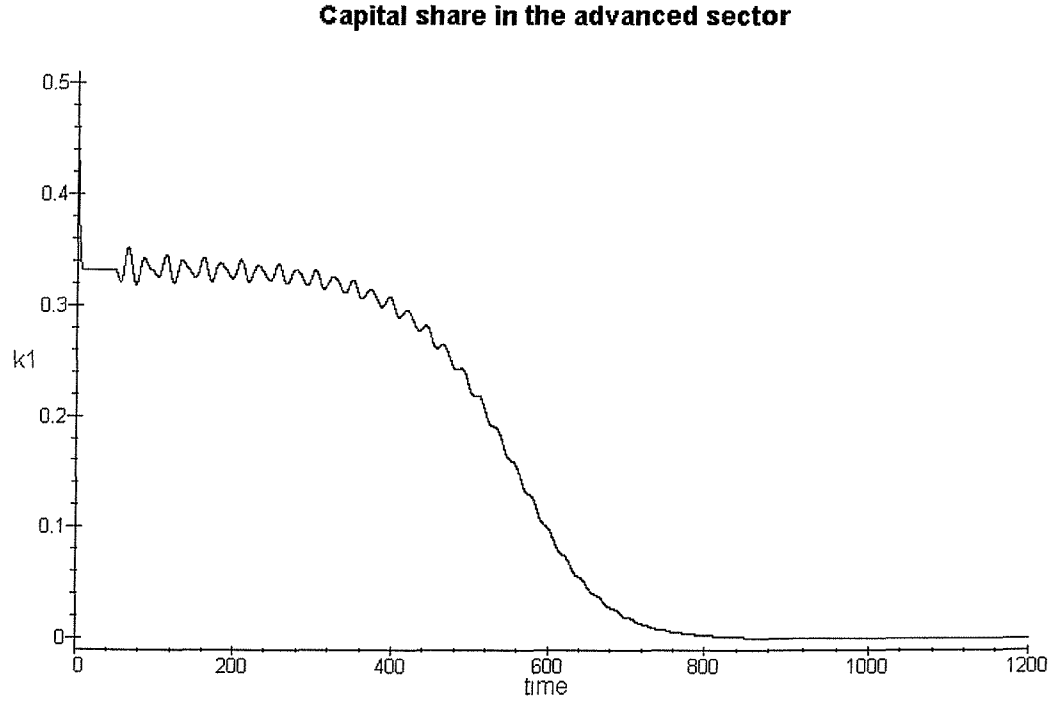

Figure 5.1

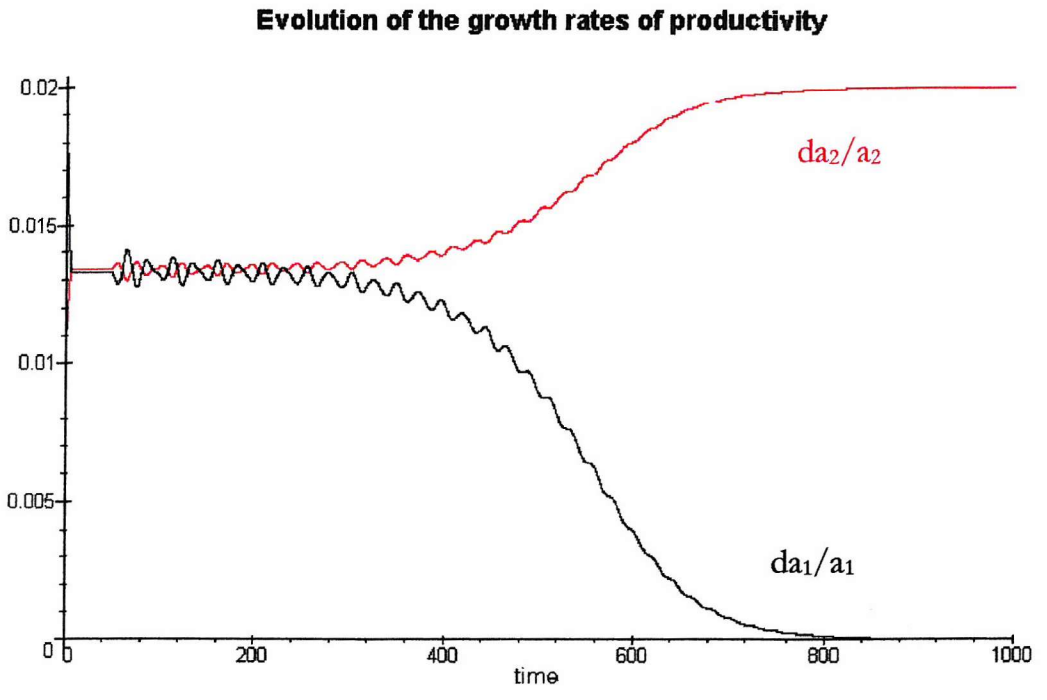

Figure 5.2

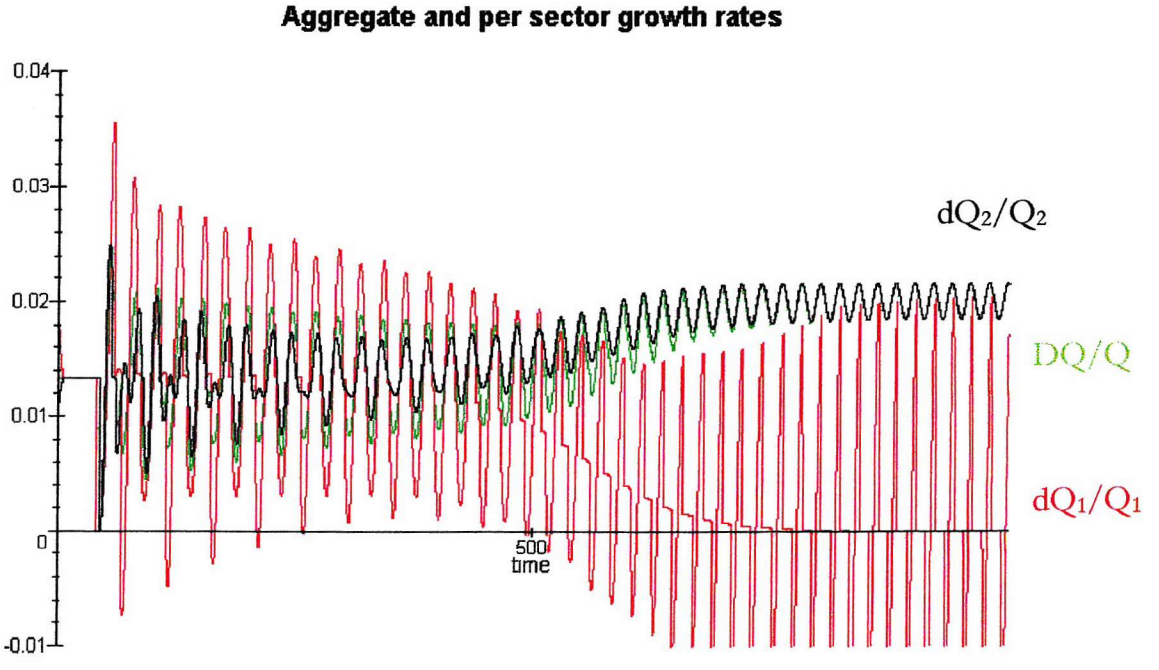

Figure 5.3

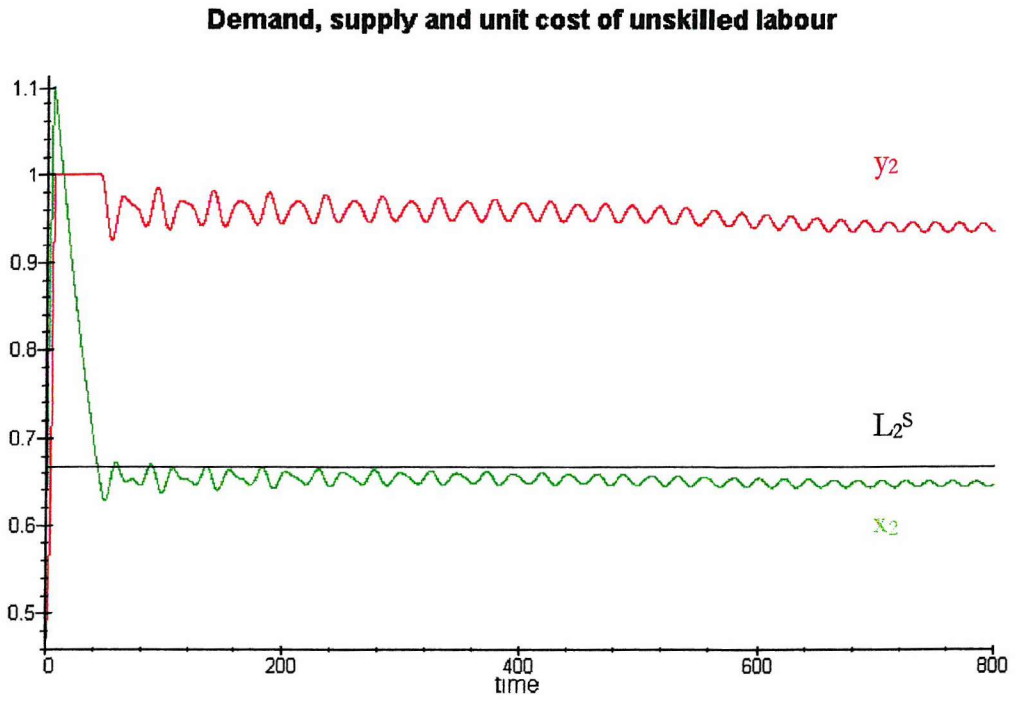

Figure 5.4

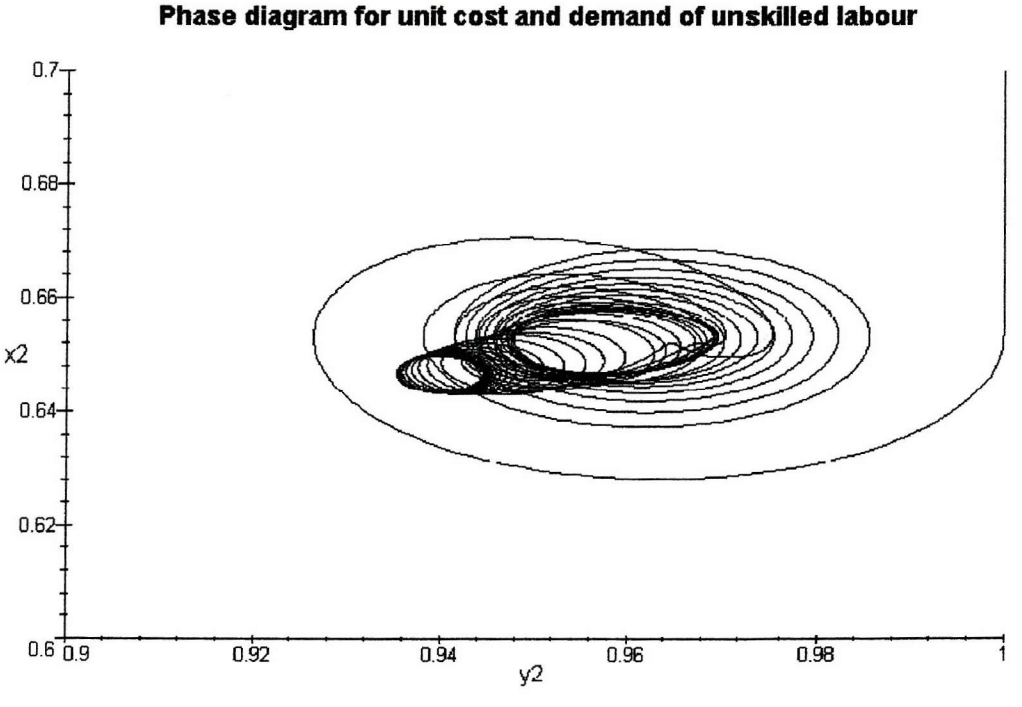

Figure 5.5

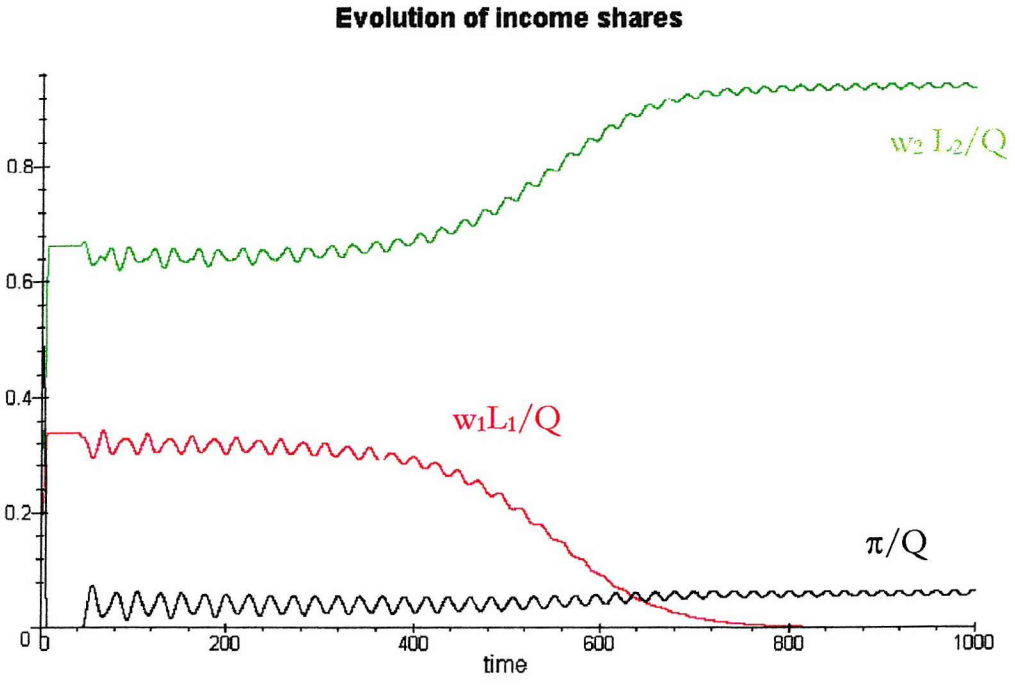

Figure 5.6

5.3.2 Second Scenario: Convergence Towards the High-Growth Steady State with Limited Skill Shortage

In this section I only slightly alter two of the parameters with respect to the previous case. This has the effect of completely upsettting the previous outcome. First, the percentage of skilled workers in the population now shifts from a third of the previous case to forty percent in the present one, thus implying a less marked skill shortage - namely s=0.4. Moreover, parameter η takes on a value less than 1, namely 0.5, indicating a less favourable distribution for labour income. All of the other values are those indicated in footnote 12.

As Figure 5.7 shows, convergence to the efficient technique now obtains. Accordingly, the analytical solution that is feasible for this setting is (A2)[15]. The individual and aggregate growth rates display the same behaviour as the previous case, but the roles are now inverted: it is the first sector that becomes the leading one,

---

[15] This solution reads as follows after having substituted for the values of the parameters: $\{\kappa_1 = 1, y_1 = 0.868,\ x_1 = 0.44,\ y_2 = 0.94,\ x_2 = 0.6\}$

whereas the second sector dwindles to a zero-growth path (Figure 5.8). One of the distinguishing features of this case is that it is now capital that is rationed, while labour reaches full employment (Fig. 5.9 and 5.10). The presence of idle capital in this case as well as of unemployment in the previous scenario, is a typical characteristic of models of the Harrodian tradition, to which the present model is similar in that it precludes marginal substitution of factors of production. Finally, it can be noticed that all of the variables do not show the peculiar cyclical behaviour observed in the previous scenario, but tend instead to converge towards a point. This is due to the change in parameter η, which implies that the type of dynamics in the leading sector of the economy is that of a stable focus rather than a centre (see discussion in section 6.5). Figure 5.11 shows indeed that the variables in the leading sector display the spiralling behaviour typical of a focus.

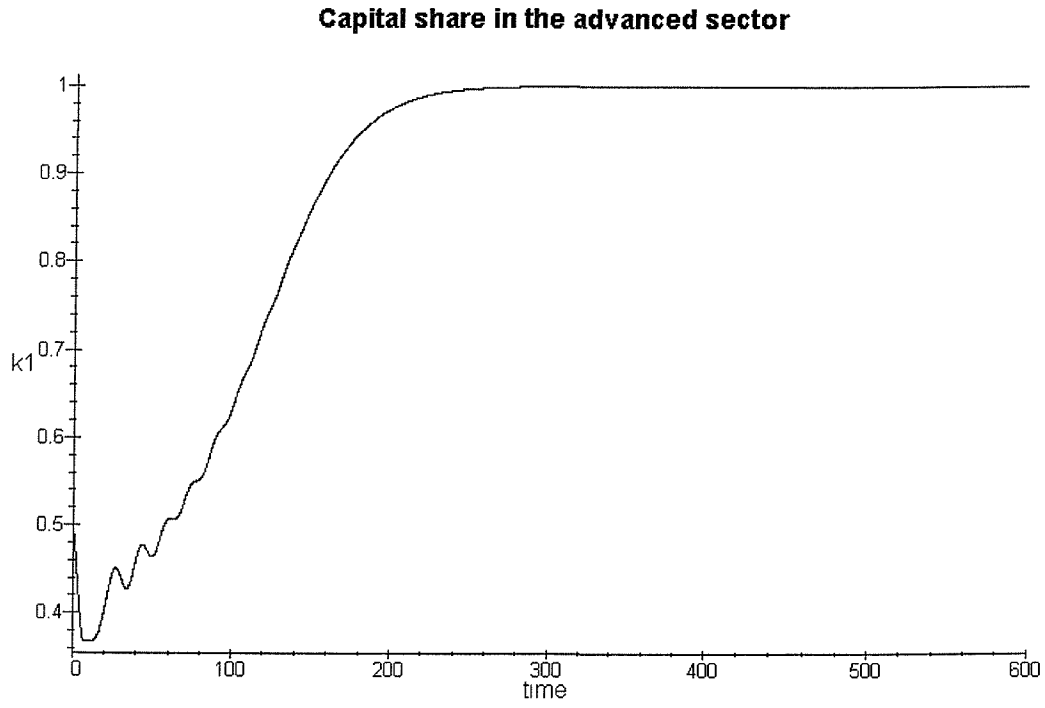

Figure 5.7

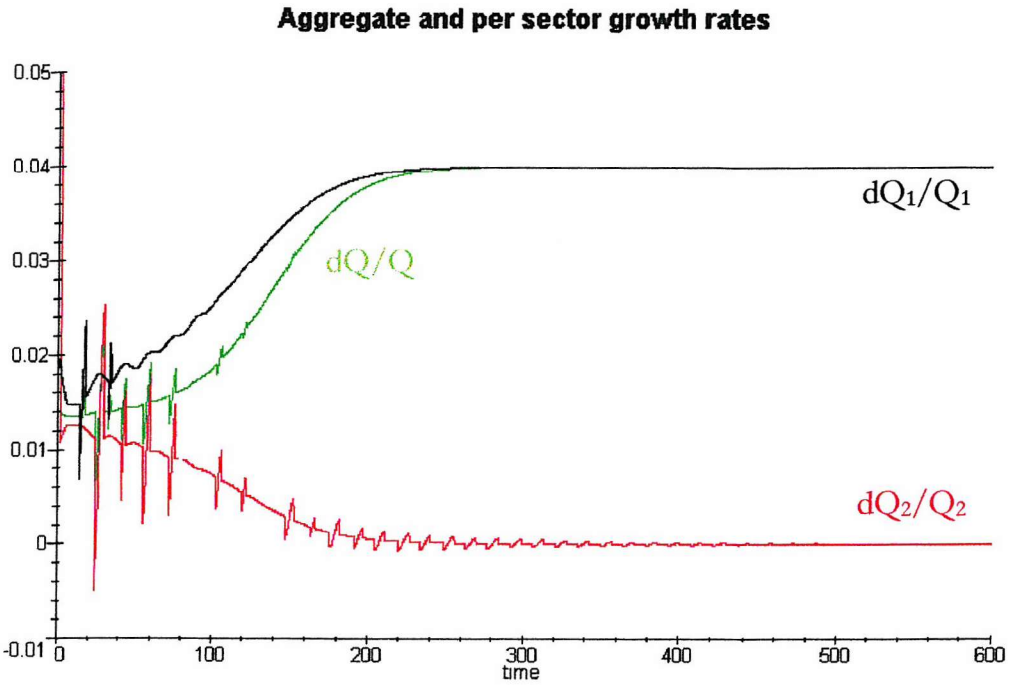

Figure 5.8

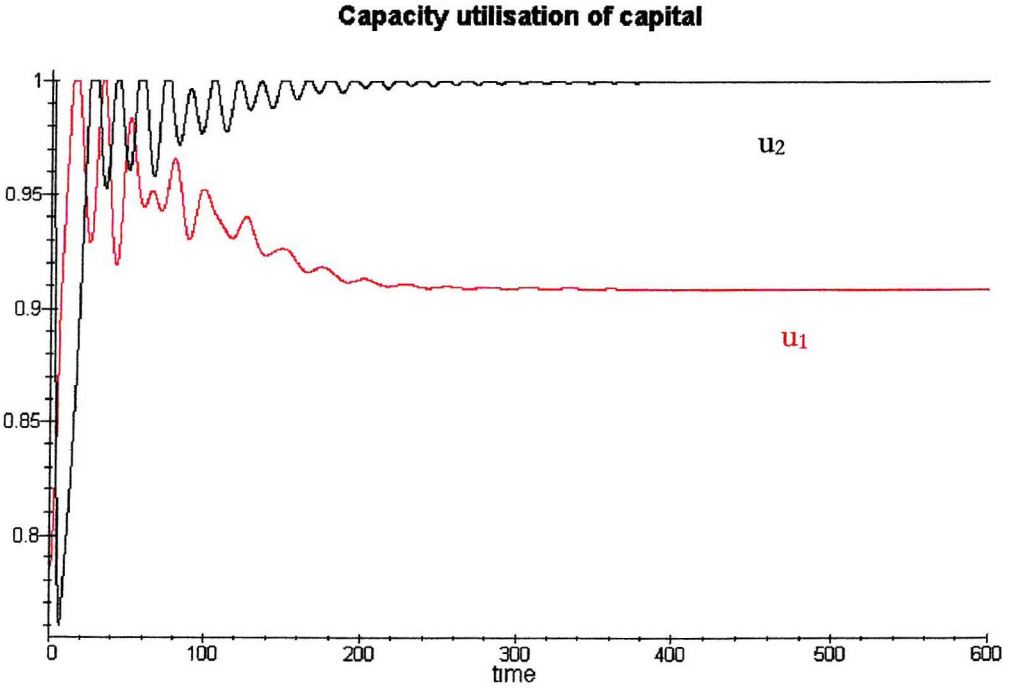

Figure 5.9

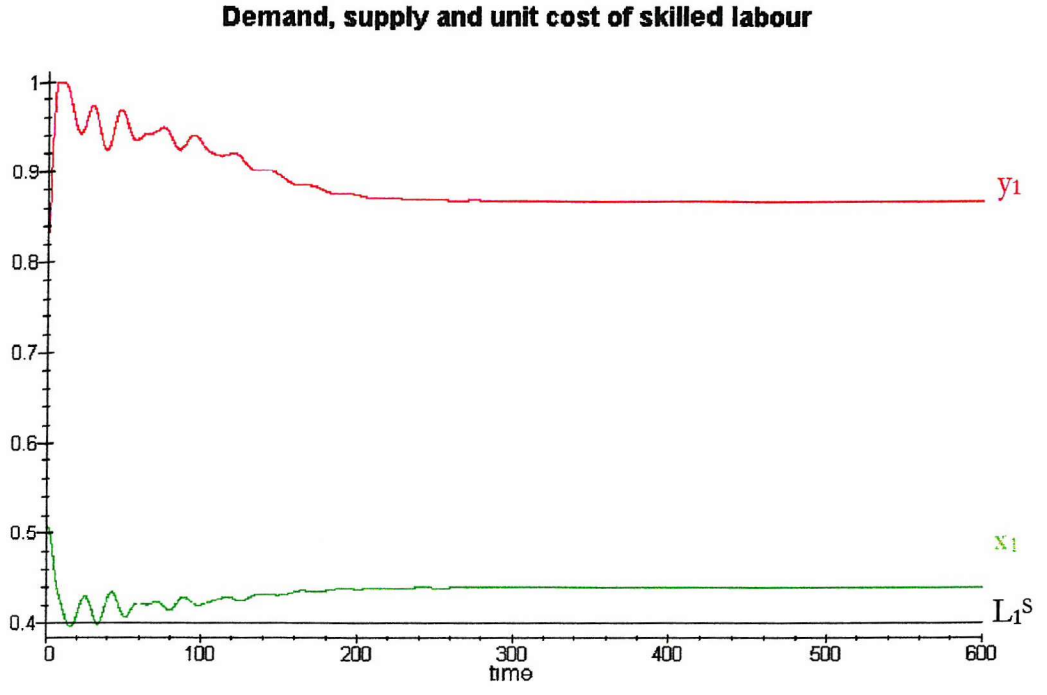

Figure 5.10

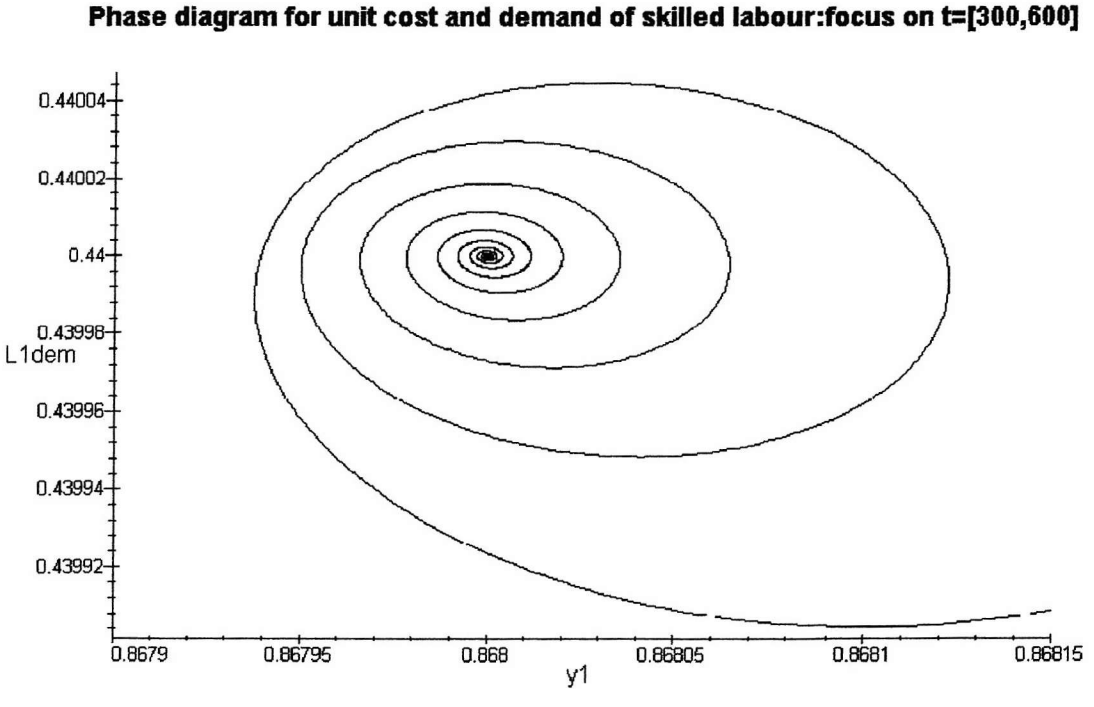

Figure 5.11

CHAPTER 6

GROWTH WITH LABOUR MOBILITY

6.1 EXTENSION OF THE MODEL

Let us now assume that workforce can be transferable from one sector of the economy to the other. Therefore, labour supply in each sector is no longer fixed, but is a function of time:

$$\begin{aligned} L^S &= L_S^S(t) + L_U^S(t) \\ L_1^S(t) &= s(t) \\ L_2^S(t) &= 1 - s(t) \end{aligned} \tag{6.1}$$

The normalisation to one of the whole of the workforce has again been adopted; furthermore, $s(t)$ denotes the percentage of the population of workers supplying labour on the skilled market as a function of time t.

Generally, the skill upgrade for a worker demands some costs due to the necessity of increasing her stock of human capital. We shall assume that the distribution of "abilities" in the population is not homogeneous, but follows a uniform distribution on the interval [0,1], where the ranking goes from the ablest individual (s=0) to the least able (s=1). The cost of upgrading is determined accordingly by means of a function called $\mu_1(s)$, which is convex and monotonically increasing:

$$\begin{aligned} &\mu_1(0) = 0, \quad \lim_{s \to 1} \mu_1(s) = +\infty \\ &\frac{d\mu_1(s)}{ds} > 0, \quad \frac{d^2\mu_1(s)}{ds^2} \geq 0 \end{aligned} \tag{6.2}$$

In this version, the cost of the upgrade is nil for the ablest worker and infinite for the least able, and the function is monotonic increasing. Notice that in this model the "ability" of a worker is not related with whether she has acquired the status of skilled, but with the ease with which such a change can be carried out.

Following an idea put forward by Blinder and Choi (1990), we can think that there also exists a "psychological" cost for the downgrading, since a worker who has started her career in the advanced sector is likely to experience a loss of social status. We can therefore assume a cost function for the downgrading symmetrical to the previous one:

$$\lim_{s\to 0}\mu_2(s)=+\infty,\ \mu_2(1)=0$$
$$\frac{d\mu_2(s)}{ds}<0,\ \frac{d^2\mu_2(s)}{ds^2}\geq 0 \qquad (6.3)$$

An example of the two costs function is given in the following diagram, which assigns a higher relevance to the "material" cost of upgrading with respect to the psychological one of downgrading:

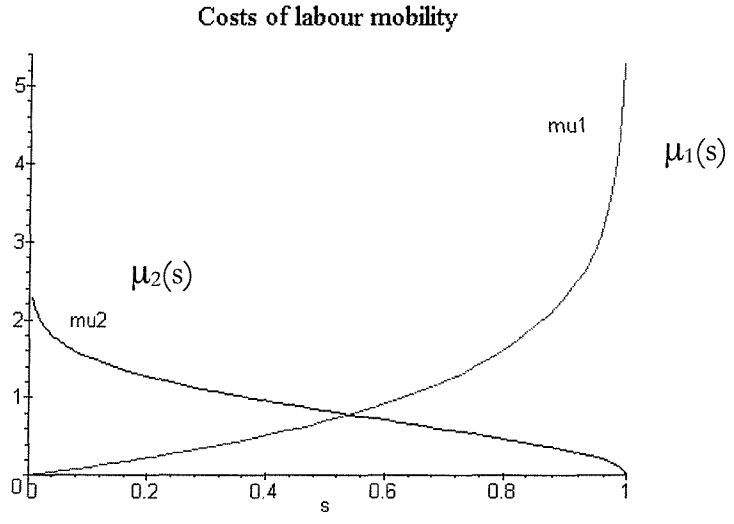

Figure 6.1

Let us assume that the utility function of the worker is linear in the wage and in the transfer cost. At each instant of time, she faces a binary decision as to whether stay in the current sector or move to the alternative one. She must then compare the expected utility gained in each sector, taking into account the mobility costs and the possibility of being unemployed. Let as assume that the probability of being left unemployed is proportional to the current unemployment rate, and that workers know

the current levels of the unemployment rates in both sectors[1]. Therefore, given that the worker with ability labelled as $\underline{s}$ is currently working in sector i, her utility function will be given by:

$$U_i^{\underline{s}}\left(w_i, w_j, L_i, L_j, L_i^S, L_j^S, t\right) = \begin{cases} \dfrac{L_i(t)}{L_i^S(t)} w_i(t) & \text{if no change} \\ \dfrac{L_j(t)}{L_j^S(t)} w_j(t) - \mu_j(\underline{s}) & \text{if change} \end{cases} \tag{6.4}$$

where, recalling the notation already introduced, L_i/L^S_i and L_j/L^S_j stand for the unemployment rates in the two sectors. Hence, a worker will decide to move to the alternative sector if the associated expected utility, net of the costs associated with the change of skill, exceeds the expected utility earned by remaining in the current sector. Even in this setting, I shall assume bounded rationality, so that information diffuses slowly and follows a process of the replicator dynamics type. Accordingly, the rate of change in the composition of the workforce is proportional to the difference between the utility earned in the two sectors, where the constant of proportionality represents the speed with which information diffuses, i.e. the probability of being selected for a random matching (see section 5.1.3). In order to avoid unnecessary complications, I assume that the first workers to move across sectors, if convenient for them to do so, are those being at the "margin" of the markets, that is, the currently least able in the case of a shift from the skilled to the unskilled labour market, or the ablest in the case of a movement in the opposite direction. Therefore, the following rule of motion obtains:

$$\dot{s} = \begin{cases} \beta s(1-s)\left[\left(\dfrac{L_1}{s}\right)w_1 - \mu_1(s) - \left(\dfrac{L_2}{1-s}\right)w_2\right] & \text{if } \left(\dfrac{L_1}{s}\right)w_1 - \mu_1(s) > \left(\dfrac{L_2}{1-s}\right)w_2 \\ -\beta s(1-s)\left[\left(\dfrac{L_2}{1-s}\right)w_2 - \mu_2(s) - \left(\dfrac{L_1}{s}\right)w_1\right] & \text{if } \left(\dfrac{L_2}{1-s}\right)w_2(t) - \mu_2(s) > \left(\dfrac{L_1}{s}\right)w_1(t) \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

[1] In this I am abstracting from the possibility that the probability of unemployment is proportional to the worker's level of ability. Also, the assumption that agents know the unemployment rate may seem to be at odds with the assumption of imperfect rationality and bounded rationality. The results we obtain, though, do not hinge upon this hypothesis.

As pointed out earlier, β measures the speed with which information is made available to agents, and their degree of rationality in processing this information is also taken into account by means of the difference in utilities.

Equation (6.5) forms a system of seven differential equations together with equations (5.25)-(5.29) and the two rules for productivity growth rates given by equation (5.2). As wages depend directly on productivity, it is no longer possible to make the system autonomous by means of the introduction of the auxiliary variable κ_1.

6.2 Analysis of Local Stability

Analysing the local stability of the steady state is now complicated by the presence of an additional equation and by the overall greater analytical complexity of the system. In order to make the derivation of some results possible at all, I first set the transfer costs μ_1 and μ_2 in equation (6.3) equal to 0, which permits a significant simplification of the expression to the following one:

$$\dot{s} = \beta s(1-s)\left[\left(\frac{L_1}{s}\right)w_1 - \left(\frac{L_2}{1-s}\right)w_2\right] \tag{6.6}$$

In fact, on the grounds of the simulations that I have conducted, such costs only appear to have a key role in determining which of the steady states will be reached, but they do not seem to affect their nature.

Introducing some other minor simplifications, we can thus find the set of steady states for the system. It would be tedious to report all of the results of the analysis. The interested reader may find more information in section 6.5. I will limit here to summarise the main results that can be drawn. The most important conclusion is that the nature of the model does not seem to be affected by the inclusion of labour mobility. In fact, the steady states found for the previous model with no mobility of labour simply 'carry over' to the present setting. More precisely, for each steady state found in the previous model there exists a steady state in the present setting whose values coincide, apart from an exception, once the steady state value for the variable s, which was a parameter in the former model, has been substituted. For instance, solution (A1) of the previous model has a 'nearly-twin' solution here:

(D1)

$$\left\{\kappa_1 = 1, \quad y_1 = 1 - cg_1, \quad x_1 = 1 - \frac{(\eta - 1)}{\gamma} g_1, \quad y_2 = \text{undetermined}, \quad x_2 = 0, \; s = 1\right\}$$

In fact, all of the steady state values in (A1) are identical to those of (D1) once $s=1$ has been substituted into the expressions of (A1). The only difference is that in the present setting y_2 is undetermined. Furthermore, the steady state corresponding to the balanced growth path (C1) is also unstable. Moreover, the extra steady states in the latter version which do not match any of the previous all turn out to be unstable.

Even in this setting, therefore, steady states are found in which there is complete allocation of the resources of the economy, be it capital or labour, into one single sector, which is the *same* for both factors. The conclusion that one could draw is thereby that market forces are sufficient in order to drive both resources to the same sector, thus avoiding the possibility of skill mismatches between the technological requirements and the workforce qualifications. Nevertheless, even in this setting they do not suffice to co-ordinate the agents on the efficient technology. In other words, the labourforce seems to behave in the same way as capital, despite the presence of switching costs that were not assumed for capital: the long-run outcome must be the complete shift of workers into the technique that becomes the leading one in the economy, the reason being that both workers and firms ultimately become attracted by the higher earnings that can be made in the leading sector of the economy. The following section reports some results of the simulations that have been conducted, focussing in particular on the role of adjustment costs on the determination of the long-run outcome.

6.3 Third and Fourth Scenarios: Skill Shortage and High vs. Low Costs of Skill Upgrading

First, we need to specify a functional form that can be used to represent the skill-upgrading costs. I shall employ a logarithmic form:

$$\mu_1(s) = \ln\left(\frac{1}{1-s}\right)^{\lambda_1} \tag{6.7}$$

λ_1 is a parameter determining the magnitude of the upgrade costs: the higher the parameter, the higher the cost for every member of the population to improve their skill.

We shall then assume a function symmetric to the previous one for the downgrading cost:

$$\mu_2(s) = \ln\left(\frac{1}{s}\right)^{\lambda_2} \tag{6.8}$$

The parameter values and initial conditions are such that the system starts with quite a marked skill shortage as in scenario 1, and with an upgrading cost relatively much higher than the downgrading cost. All of the other conditions denote again a situation of initial symmetry between the two sectors[2]. The main upshot is that even in this case a result of convergence towards the slow-growth steady state obtains, which portrays a dynamics for the distribution of investments across sectors quite similar to that of the first scenario (Fig. 5.1). The underlying causes are the same as those stressed for the previous ones. In addition, productivity affects wages in such a way that workers are attracted to what soon becomes the leading sector of the economy, as shown by figure 4.2. The picture shows how the flow of labour is related to the wages differential. Therefore, productivity growth acts as the main factor to attract resources allocation both for capital and labour. According to the simulations conducted, this seems to be a general result, so that one could conclude that there is no risk of mismatch between labour demand and supply within this model. That is, a situation where capital and labour are allocated in the two different sector[3], is avoided. However, what cannot in general be impeded is, again, a lock-in effect to the inefficient technique.

An interesting feature of the model is its sensitiveness to the structure of the economy, and in particular to the magnitude of the skill upgrade costs. Indeed, it is sufficient to slightly shift the value of the related parameters to reverse the result of

---

[2] In particular: $\{\alpha:=1; c:=3; \gamma_1:=1; \gamma_2=1; \beta=1; \eta:=1.5; g1:=0.04; g2:=0.02; \lambda_1=1, \lambda_2= 0.1\}$ $\{y_1(0)=0.5; y_2(0)=0.5; x_1(0)=.5; x_2(0)=.5; a_1(0)=1; a_2(0)=1; s(0)=1/3\}$.

[3] In fact, many of the additional steady states that can be found in the labour mobility version of the model display firms entirely concentrated in one sector of the economy and labour in the other. However, these steady states turn out to be unstable from a preliminary analysis.

convergence towards the inefficient technique. In fact, after considering a set of parameters identical to the previous one but for the parameter λ_1, which is now shifted to a lower value (0.975 instead of 1) denoting a lesser transfer cost for skill upgrading, we obtain a result of convergence towards the first sector, analogous to that obtained in the second scenario (Figure 3.7).

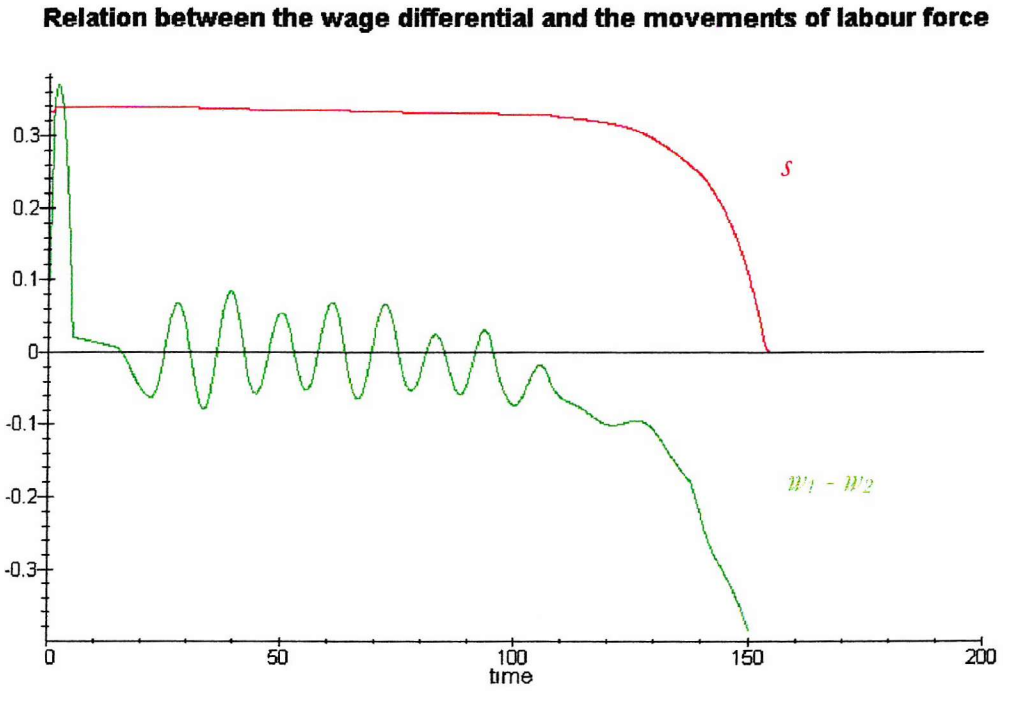

Figure 6.2

6.4 CONCLUSIONS

A model of growth with multiple steady states has been developed, which depicts the evolution of an economy characterised by localized technical change, i.e. spillovers taking place only at the technique-specific level, bounded rationality, which determine a replicator-type dynamics for aggregate behaviour of agents, and non-instantaneous market clearing. On the grounds of the local stability analysis and the simulations conducted, a high and a low-growth steady states, in which all factors are allocated to the skilled and unskilled-intensive technique respectively, obtain as stable equilibria of the model. A balanced growth path steady state, characterised by both sectors growing at the same rate, turns out to be unstable, because of the cumulative process of technical change. Some structural conditions determining the outcome of convergence have been highlighted, such as the degree of skill shortage in the case of non-transferability of labour, and the extent of skill upgrading costs in the case of labour mobility.

The low-growth steady state may be interpreted as the result of a co-ordination failure amongst the variety of agents making up the economy, i.e. workers and firms: both outcomes in which all agents converge towards the same sector are indeed 'equilibria' of the interaction, but the co-ordination on the high-growth equilibrium Pareto-dominates the other. In other words, in this world of gradual adjustment towards equilibrium and optimality and slow diffusion of information, market forces suffice to impede the mismatch of productive factors, but they do not always provide enough incentives to converge towards the efficient outcome. This in particular is the case when adverse initial structural conditions for the economy occur.

The analysis conducted has some straightforward, but significant, implications of political economy. In fact, a policy of training the unskilled work force, softening the initial skill shortage and lowering the skill upgrade costs, would make it possible to overcome the sub-optimal outcome. However, since the transition from the inefficient to the Pareto-efficient outcome requires some groups to give up part of their income shares in order to pay for the costs of the policy, in exchange of a benefit in the future,

then some form of intertemporal agreement between the parties is necessary to guarantee the undertaking of the plan. As we all know, this is far from an easy requirement, though, especially in less developed countries where institutions are typically rather unstable. In more general terms, the paper stresses the complexity of a process of catching-up in presence of adverse structural conditions: even when a potentially more efficient technology is available in an economy, a lack of skill by the agents, concerning both workers as to their capacity to adapt to that technology, and firms as to their ability to exploit it, may thwart the economic incentives necessary to undertake the high-growth path.

6.5 APPENDIX

6.5.1 Analysis of the Steady States for the Case of Non-Mobility of Labour

A1) $$\left\{\kappa_1 = 1 \quad y_1 = 1 - cg_1 \quad x_1 = L_1^S - \frac{(\eta - 1)}{\gamma} g_1 \quad y_2 = \frac{1 + \alpha(1 - cg_1)}{1 + \alpha} \quad x_2 = L_2^S\right\}$$

The stability analysis is complicated by the fact that the variable x_2 is located just on the threshold level where the related equation (5.28) changes its expression. This implies that the Jacobian on the right neighbourhood of the point differs from that of the left neighbourhood. On the left neighbourhood of $x_2 = L_2^S$, where both u_1 and u_2 equals 1, we get the following set of eigenvalues:

$$\left\{-g_1, \frac{\sqrt{c\left[\gamma L_1^S - g_1(\eta - 1)\right](1 - cg_1)}}{c} i, \quad -\frac{\sqrt{c\left[\gamma L_1^S - g_1(\eta - 1)\right](1 - cg_1)}}{c} i, \right.$$
$$\left. \frac{\sqrt{c\gamma L_2^S(1 + \alpha(1 - cg_1))}}{c} i, \frac{\sqrt{c\gamma L_2^S(1 + \alpha(1 - cg_1))}}{c} i\right\} \qquad (6.9)$$

Intuitively, the negative eigenvalue can be associated with the κ_1 axis, indicating the profitability of allocating capital in sector 1. Furthermore, two pairs of purely imaginary eigenvalues are obtained. Even though their sign is dubious, for the range of parameters economically meaningful we can be sure the argument of the square root is actually negative. In fact, g_i must have an order of magnitude at least 10 times smallest than the value of the other parameters, and this ensures in particular that

$$1 - cg_i < 1 \qquad (6.10)$$

From linear stability theory we know that a two-equation linear system having a couple of purely imaginary eigenvalues is a centre. However, this conclusion does not necessarily carry over to non-linear systems. If this was the case, nevertheless, we may think that each couple of eigenvalues actually describes the dynamics of each of the two sectors, so that labour demand and wages should display the cyclical behaviour typical of a centre.

In the right neighbourhood of x_2, when $u_2<1$, we find the following set of eigenvalues:

$$\left\{-g_1,\ \frac{\sqrt{c\left[\gamma L_1^S-g_1(\eta-1)\right](1-cg_1)}}{c}i,\quad -\frac{\sqrt{c\left[\gamma L_1^S-g_1(\eta-1)\right](1-cg_1)}}{c}i,\right.$$
$$\left.-\frac{\alpha g_1}{2}-\frac{\sqrt{c\left[4\gamma L_2^S(1+\alpha(1-cg_1))-c\alpha^2 g_1^2\right]}}{c}i,\quad -\frac{\alpha g_1}{2}+\frac{\sqrt{c\left[4\gamma L_2^S(1+\alpha(1-cg_1))-c\alpha^2 g_1^2\right]}}{c}i,\right\} \tag{6.11}$$

It is notable that the negative eigenvalue and one of the two couples of imaginary eigenvalues coincide with what found for the left neighbourhood. Conversely, we now have a couple of complex conjugate eigenvalues with negative real part, whose dynamics would then be that of a stable focus generating trajectories that converge spiralling to a point.

We can indeed be sure, by means of analytical considerations, that the couple of purely imaginary eigenvalues that remain unchanged in the two neighbourhoods can be associated with the first sector. In fact, looking at the system of differential equations (5.25)-(5.29), one can notice that once κ_7 is equal to 0, as is the case asymptotically, then the second sector becomes "autonomous" from the variables of the first sector, thus assuming the form of a Lotka-Volterra two-equation system. Hence, the variables of the second sector must asymptotically behave like a centre, and converge towards a limit cycle. That this is the case can be derived from picture 3.5.

As for the first sector, the presence of different pairs of eigenvalues in the two neighbourhoods of the solution, which would generate different dynamical behaviour if taken singularly, makes it impossible to state their dynamical behaviour with certainty. The most likely conjecture, also supported by some further evidence derived from graphical analysis not reported here, is that variables of the first sector oscillate on a torus, although possible behaviours would also be those of a limit cycle with a

slower process of convergence than the second sector, and a strange attractor, i.e. a behaviour characterised by chaotic evolution within a limited manifold.

$$(A2)\quad \left\{\kappa_1 = 1,\ y_1 = 1 - cg_1 - c\left(\frac{1-\eta}{\gamma L_1^S}\right)g_1^2,\ x_1 = L_1^S + \frac{(1-\eta)}{\gamma}g_1,\ y_2 = \frac{1+\alpha(1-cg_1)}{1+\alpha},\ x_2 = L_2^S\right\}$$

The properties of stability of the point must again be conducted considering two different sets of eigenvalues. In the left neighbourhood of $x_2 = L_2^S$ we shall observe:

$$\left\{-g_1,\ -\frac{g_1}{2} - \frac{\sqrt{c\left[4\gamma L_1^S(1-cg_1) - cg_1^2(5-4\eta)\right]}}{c}i,\quad -\frac{g_1}{2} + \frac{\sqrt{c\left[4\gamma L_1^S(1-cg_1) - cg_1^2(5-4\eta)\right]}}{c}i,\right.$$
$$\left.\frac{\sqrt{c\gamma L_2^S(1+\alpha(1-cg_1))}}{c}i,\ \frac{\sqrt{c\gamma L_2^S(1+\alpha(1-cg_1))}}{c}i\right\} \tag{6.12}$$

In the right neighbourhood the eigenvalues are as follows:

$$\left\{-g_1,\ -\frac{g_1}{2} - \frac{\sqrt{c\left[4\gamma L_1^S(1-cg_1) - g_1^2c(5-4\eta)\right]}}{c}i,\quad -\frac{g_1}{2} + \frac{\sqrt{c\left[4\gamma L_1^S(1-cg_1) - g_1^2c(5-4\eta)\right]}}{c}i,\right.$$
$$\left.-\frac{\alpha g_1}{2} - \frac{\sqrt{c\left[4\gamma L_2^S(1+\alpha(1-cg_1)) - c\alpha^2 g_1^2\right]}}{c}i,\quad -\frac{\alpha g_1}{2} + \frac{\sqrt{c\left[4\gamma L_2^S(1+\alpha(1-cg_1)) - c\alpha^2 g_1^2\right]}}{c}i,\right\} \tag{6.13}$$

Hence, even in this case we have three eigenvalues that remain the same in both neighbourhoods, which are possibly associated with the κ_1 axis and with the variables of the leading sector. However, instead of having two couples of imaginary values we find a pair of complex conjugates eigenvalues: this leads us to think that the variables in the first sector converge spiralling to a focus, as displayed in Fig. 3.11. As in the previous case, we have a couple of imaginary eigenvalues on one side and a pair of stable complex conjugates on the other for the residual sector: this is a dynamic behaviour not easily classifiable, alike that found for solution (A1).

I will not deal with the analysis of steady states (B1) and (B2) presented in section 5.2.2.B since they are symmetric to solutions A1 and A2.

$$(C1)\quad \left\{\begin{array}{lll} \kappa_1 = \dfrac{g_2}{g_1+g_2} & y_1 = \dfrac{g_2 + g_1(1-cg_2)}{g_1+g_2} & x_1 = L_1^S - \dfrac{(\eta-1)}{\gamma}g_1g_2 \\ y_2 = \dfrac{g_2 + g_1(1-cg_2)}{g_1+g_2} & x_2 = L_2^S - \dfrac{(\eta-1)}{\gamma}g_1g_2 & \end{array}\right\}$$

The analytical expression of the set of eigenvalues for steady state (C1) relative to a balanced growth path is rather complicated and will not be reported. For an

economically reasonable set of parameters, however, one obtains the following set of values, where I is the purely imaginary unit:

$$\{-.0165511798 + .5688615151\,I,\ .01356157422,\ -.0165511798 - .5688615151\,I,\ -.0102296072 + .3609416065\,I,\ -.0102296072 - .3609416065\,I\}$$

Intuitively, we can associate the 4 complex eigenvalues that can be found to the variables related to each sector –labour demand and labour unit cost. The positive eigenvalue can instead be associated with the κ_1 co-ordinate, on the grounds of economic consideration set out in section 5.2.1.

6.5.2 Analysis of steady states for the case of labour mobility

As already pointed out, the analysis of local stability for the case of labour mobility reveals the close similarity from the economic standpoint between the steady states found in the two cases. However, the reader interested in the mathematical details may want to explore the peculiarities of this, more complex, case. In this section I thereby offer a brief summary of the results obtained.

$$\text{D1)}\ \left\{\kappa_1 = 1 \quad y_1 = 1 - cg_1 \quad x_1 = 1 - \frac{(\eta - 1)}{\gamma} g_1 \quad y_2 = \text{undetermined} \quad x_2 = 0 \quad s = 1\right\}$$

I have already noticed how clearly does this solution correspond with solution (A1) in that, apart from y_2 being now undetermined, (A1) boils down to (D1) once the steady state value $s=1$ is substituted into the steady states values of the other variables. Since variable x_2 is located on the edge of its admissible values, it obviously suffices to find eigenvalues only on the relevant neighbourhood of the space. The following set obtains:

$$\left\{\frac{(1+\alpha)(1-y_2)-\alpha c g_1}{c}, \frac{(1+\alpha)(1-y_2-cg_1)}{c}, \frac{\sqrt{c[\gamma - g_1(\eta-1)][(1-cg_1)}}{c} i, \quad -\frac{\sqrt{c[\gamma 1 - g_1(\eta-1)][(1-cg_1)}}{c} i,\right.$$

$$\left. -\frac{a_1\beta}{\gamma}[\gamma - g_1(\eta-1)][(1-cg_1),\ 0\right\} \tag{6.14}$$

No general conclusion can in general be drawn on the sign of the eigenvalues. However, some speculative considerations can be put forward. First, let us compare this with the set of eigenvalues found for (A1): two of the purely imaginary eigenvalues coincide, once the value of s has been substituted; presumably, they are

associated with the Lotka-Volterra dynamics setting in within the leading sector of the economy. The first two eigenvalues of (D1) have now a dubious sign. However, for a significant set of the parameters, and for y_2 sufficiently close to 1, which on the grounds of the simulations conducted seems indeed to be the case, their values turn out to be negative. We finally have one negative eigenvalue, again for a realistic value of the parameters, and one equal to zero. Overall, therefore, this analysis cannot be conducive to any definite conclusion, because of the presence of eigenvalues with real part equal to nil. However, the simulations conducted prove indeed that this steady state turns out to be an attractor of the system, the two variables associated with the leading sector, x_1 and y_1 that is, moving along a close orbit whose centre is that indicated in (D1), y_2 converging to 1, and all of the other variables converging to the values prescribed in (D1).

$$\text{(D2)}\quad \left\{\kappa_1 = 1,\ y_1 = 1 - cg_1 - c\left(\frac{1-\eta}{\gamma}\right)g_1^2,\ x_1 = 1 + \frac{(1-\eta)}{\gamma}g_1,\ y_2 = \text{undetermined},\ x_2 = 0,\ s = 0\right\}$$

Alike (D1), solution (D2), which holds under the constraints that η is less than 1, has a corresponding solution in (A2) in that the latter obtains if one substitutes $s=1$ for the parameter s in (A2). The only difference lies in that y_2 is undetermined in (D2), whereas it takes a definite expression in (A2). Assigning for simplicity $y_2 = 1$, which is the most likely value assumed by this variable in the residual sector of the economy, and the one actually observed in the simulations that have been conducted, the following set of eigenvalues obtains:

$$\left\{-\frac{g_1}{2} - \frac{\sqrt{c\left[4\gamma L_1^S(1-cg_1) - cg_1^2(5-4\eta)\right]}}{c}i,\quad -\frac{g_1}{2} + \frac{\sqrt{c\left[4\gamma L_1^S(1-cg_1) - cg_1^2(5-4\eta)\right]}}{c}i,\right. \tag{6.15}$$

$$\left. -\alpha g_1,\ -g_1(1+\alpha),\ 0,\ -\beta a_1\left[(1-cg_1) - cg_1^2\left(\frac{1-\eta}{\gamma}\right)\right] + \beta a_2\right\}$$

Interesting analogies with solution (A2) can be found here, too. The first pair of complex conjugates eigenvalues is identical to that found for (A2): this is likely to be associated with the dynamics in the leading sector of the economy, and indeed in the simulations I have conducted variables x_1 and y_1 show the typical behaviour of a focus. The other eigenvalues are either negative – in particular the latter is certainly negative as a_1 in the long run outstrips a_2 by a large amount – or equal to 0, which is likely to be

associated with y_2. That this solution is indeed an attractor for the system has been verified through numerical simulations.

Finally, symmetric solutions to the pair now illustrated, characterised by convergence to the second sector, can be found. These can be associated with solutions (B1) and (B2) for the same reasons set out above:

(E1) $\left\{\kappa_1 = 0 \quad y_2 = 1 - cg_2 \quad x_2 = 1 - \frac{(\eta - 1)}{\gamma} g_2 \quad y_1 = \text{undetermined} \quad x_1 = 0 \quad s = 0\right\}$

(E2)

$$\left\{\kappa_1 = 0,\ y_2 = 1 - cg_2 - c\left(\frac{1-\eta}{\gamma}\right)g_2^2,\ x_2 = 1 + \frac{(1-\eta)}{\gamma} g_2,\ y_1 = \text{undetermined},\ x_1 = 0, s = 0\right\}$$

Not surprisingly, these steady states show symmetric properties of stability to those just examined.

Finally, an equivalent solution to he balanced growth path solution (C1) seems to obtain even in this case, even though a complete analytical solution is not possible. However, after having assigned the numerical values used in the previous simulations to the parameters, one can express all the variables as a function of the particular value taken on by s. This solution looks indeed the exact analogous of solution (C1):

$$\left\{\begin{array}{lll} \kappa_1 = \dfrac{g_2}{g_1 + g_2} & y_1 = \dfrac{g_2 + g_1(1 - cg_2)}{g_1 + g_2} & x_1 = s - \dfrac{(\eta - 1)}{\gamma} g_1 g_2 \\ y_2 = \dfrac{g_2 + g_1(1 - cg_2)}{g_1 + g_2} & x_2 = (1 - s) - \dfrac{(\eta - 1)}{\gamma} g_1 g_2 & s = \dfrac{1199999998}{2399999998} \approx 0.499 \end{array}\right\}$$

Not surprisingly, this turns out to be unstable.

THIRD PART
TECHNOLOGY AND SOCIAL NORMS

INTRODUCTION

In this final Chapter, I would like to return to the issue identified at the outset (see Preface of the thesis), and investigate the idea of the mutual interaction between social norms and economic performance. In particular, building on the material developed in the earlier Parts of the thesis, I aim to model the relationship between technology and social norms in such a way as to capture both the fact that the type of technology adopted in an economy may influence its institutional structures and the pattern of norms, and the fact that particular social norms may in turn affect the adoption of specific technologies.

In particular, I try to model the relationship between the economic and the social side as twofold: on the one hand, the type of technology generally adopted in an economic system may call for particular types of institutions and influence the established social norms. In particular, economic activities with different degrees of risk and different patterns of distribution of the costs of risk across agents may call for different types of social institutions. Ideally, an optimal institution should alleviate the bad consequences and stimulate the positive aspects linked with the undertaking of risk and competition within an economic system[1]. On the other hand, norms may foster or thwart the adoption of some rather than other technologies. For instance, a

---

[1] To be sure, the credit system is the institution than in the reality of market economies permits to handle risk at the aggregate level. Although the model developed in this chapter is characterised by such a level of abstraction as to make it impossible its explicit consideration, one could read the main results of the model as related with the type and degree of development of the credit system. However, it has to be noticed that credit is only one of the many ways in which a modern society manages risk.

'collectivist' type of norm, as opposed to one that elicited the allocation of resource on an 'individualistic' basis (see Greif (1994)), may be at variance with a technology in which individual contributions were substitutes. The opposite would of course be true with respect to a production function that required complementary individual contributions.

In the model developed in this Chapter some of these issues will be addressed by portraying the strategic relationship between two entrepreneurs involved in a 'race for the market'. Building on the Grossman-Hart's framework for productive integration (1986), agents will have two strategies available in *handling* risk: they could either select a risk-sharing economic activity, which requires an investment in 'co-operative skills' and is carried out through a joint production function; or, they can opt for a more risky activity, which is moulded as a winner-takes-all contest. Social norms are shaped as regularities of behaviour associated with system of expectations, which a normative prompt provide to individual motivations along the lines of the models proposed in the first part of the thesis. The associated sets of institutions could then be called *co-operative* or *competitive* according to which type of technology and social norms emerge. Social norms will be seen as having two potentially opposite roles: on the one hand they could be the main determinant for setting up efficient co-operative economic activities. In this sense, building on the evolutionary approach that will be adopted in the interpretation of the model, their role could be depicted as one of providing a 'social memory' of past beneficial interactions, which constrain present interactions through the normative power of expectations and perceptions of the public interest. On the other hand, norms could also take on a 'conservative' role in that they could hamper the shift to alternative better technologies, or the activity of experimentation and innovation, when social norms sustain the practise of engaging in co-operative economic activities. This may originate a situation of *mismatch* between technologies and institutions, analogous to the instance of slow-growth trap described in the second part of the thesis.

Some final caveats are needed before starting off. First, one could wonder if it is sensible to talk about a technology as a whole for the economic system. That this is indeed the case can be argued by referring to the work of some economists (e.g. Dosi

et al. (1988)) have introduced the concept of 'national technological systems', and one can always hope to identify the 'leading' as opposed to 'residual' sectors in an economy, thus focussing on the key ones and neglecting the others (see previous chapter on growth). For 'type' of technology, here, I refer in particular to the *optimal scale* of the production function, which may favour enterprises with high fixed costs rather than those with no or small fixed costs. In particular, the key property is the super-additivity of the cost function, which will be partly captured in equations 7.3 and 7.4. Although I will not further this argument in the remained of the Chapter, the main idea is that the higher the amount of investment that is needed, the higher the 'cost' of risk, as clearly the losses in case of unsuccessful business is higher. Of course, not only are *static* returns to scale relevant, but also *dynamic* returns to scale, which refers to the gains in productivity stemming from learning by doing, learning to learn, innovations and all the related issues (appropriability Vs imitation), are of remarkable importance.

Second, in the light of the typical economic approach of methodological individualism, the situation of interaction described in the model should be understood as a stylised relationship that is representative, subject to minor changes, of a variety of economic relationships within a society. For instance, the same model could also illustrate an entrepreneur-employee relationship as to the mutual investment to be carried out within the joint productive activity; or that between two employees as to the 'insurance' against unemployment. Therefore, I see the model as being descriptive of some basic relationship characterising contemporary societies, thus its results could be of some significance in the study of how a society copes with industrial relations, technological research, and systems of welfare.

Having said that, though, let me be clear that the purpose of the model is to *describe*, rather than to *explain*, the sort of general arguments so far illustrated. In fact, given the high degree of abstraction that is needed in order to ground social institutions on individual behaviour, it would be pretentious to draw direct implications from the straight results of such a simple model, without the guard of a more careful consideration of the array of factors that are inevitably left aside. Therefore, one should read the model as illustrative of *some* of the relationships that, in

my view, occur between technological systems and social norms, rather than as a thorough account of social institutions. In particular, some readers have been misled by the apparent *technological determinism* that seems to drive the model; conversely, some others have been critical on the introduction of other-regarding motivations in the second part of the chapter as a rather *ad hoc* assumption, which simply begs the question of how co-operative behaviour emerges within a society. To prevent this type of criticisms, though, I will not push the argument any further than arguing that some types of social norms are more *likely* to emerge within particular technological framework, or that the existence of some *correlations* between social institutions and technological systems can in fact be supported on theoretical grounds.

CHAPTER 7

A MODEL OF NORMS SELECTION AND ECONOMIC PERFORMANCE THROUGH INSTITUTIONAL GOVERNANCE OF UNCERTAINTY

7.1 THE SETTING OF THE MODEL

7.1.1 The Timing of the Model

The model that I will study in this chapter can be traced back to the strand of literature originated by Grossman and Hart's (GH in the following) classic account of vertical and lateral integration (1986); in fact, in my model the succession of players and 'Nature"s actions is like that of GH, with a first stage of *commitments* followed by a Nature's action, and then another action from players. My model, though, further includes another move from Nature in one version of the final stage. Moreover, the usual hold-up problem will be seen to be a crucial occurrence of my model, too. However, my focus will be on how the different technological conditions determining *vertical* integration as opposed to *non-integration* play a key role in the players' choice. Also, the lack of complete information will be a crucial determinant of the outcome of the interaction. With respect to this point, I will take on the usual Knightian distinction between *risk* and *uncertainty*, intended as *measurable* and *non-measurable*, respectively, handling of situations of *lack of information*. Due to the analytical complexity of the former aspect, I will be content with developing a formal analysis of only the latter aspect, relying on some qualitative considerations as far as risk is concerned.

The game as a whole may be conceived as one of a *potential* 'race for the market' between two entrepreneurs. This is only 'potential' as they in fact have a choice between actually going for the race in a *winner-takes-all* type of contest, or settling down

to establish a *joint* enterprise, thus splitting half of the stake in play. The stake consists of the demand for the product in a market - from which the phrase 'race for the market' is derived - so that the player who turns out to be the winner in the contest ensures for herself the whole demand and leaves nothing to the other; conversely, if they choose not to race against each other, the two players will share the profits deriving from the joint enterprise.

The final decision as to whether race or not is affected by what happens in the first two stages of the game, where agents determine their degree of commitment to a would-be joint enterprise through their amount of effort invested in *co-operative* as opposed to *competitive* skills. This is determined by the variable α_l [2], which indicates the *share* of investment in human capital devoted to the development of 'co-operative' skills. The investments in the two types of skills are supposed to be incompatible, so that investing an additional unit of resources in co-operative skill implies a reduction in the investment in the alternative type of skill.

Once the investment phase has been completed, 'Nature' has the role of determining individual productivities, which are represented by the pair of variables (θ_l, ν_l). They represent, respectively, the *individual* contribution to productivity in the *co-operative* enterprise, and individual productivity in the *competitive* enterprises, respectively. That is, the higher θ_l, the higher the expected payoff from *joint* production; the higher ν_l, the higher the expected payoff in *individual* productivity. Or, to be more precise, if individual productivity is higher, then the expected payoff of the corresponding productive activity *is not smaller*. For simplicity, I assume that the determination of individual productivities follows a random binomial distribution, so that there are two only possible outcomes available: 0 and 1, which correspond with low and high individual productivity respectively. The investment effected in the previous stage influence the realisation of productivities: in fact, the higher the investment in a type of skill, the higher the probability of obtaining high productivity in the related skill. For instance, an increase in α_l will bring about at the same time an

---

[2] As before, I shall label each agent with the letter i and j, and refer to a generic player between the two as l, with $l=\{i, j\}$.

increase in the probability of receiving a high value of θ_i, and a *decrease* in the probability of receiving a high realisation of v_i.

In particular, overall productivity within the joint production function can be generally thought of as being determined by the degree of complementarities between individual inputs. For simplicity, I study the two most extreme cases that can occur: perfect substitutability and perfect complementarity between individual inputs, which, as will be shown, lead to very different strategic situations[3]. Hence, if an individual has obtained a high θ_i, she will be not sure that joint productivity will also be high: this will only occur if her partner has got high productivity as well, or if individual inputs into joint production have high degrees of substitutability. Productivity in individual competitive production functions is instead directly determined by the related individual productivities. Moreover, it is also assumed that v_i also decides the winner of the race: in a very simple way, the winner will be the player with a higher value of v_i, and cases of draw will be resolved through the toss of a fair coin[4].

In this first version of the model, I assume that agents are risk-neutral; section 7.4, instead, offers an insight into the cases of risk-aversion of agents and radical uncertainty. For obvious reasons, only the first case has been analysed in quantitative terms. I have to stress from the outset that the way productive activities are modelled makes the competitive activity generally more risky, in the sense that its ex-ante variance is generally higher than the other. This is clearly recognisable if one consider that in the case the race for the market is lost, the loser gets a payoff of zero, whereas in the co-operative case the payoff in the 'bad' states of the world, i.e. when individual contributions to joint productivity are low, is nonetheless positive. This is enough to make the competitive activity more risky, as, for most values of the parameters, the payoffs difference in the good states, i.e. when winning the race in the competitive

---

[3] The case of complementarity between individual inputs could also be thought of as arising from the possibility of a 'bad match' in the competencies of the two partners, within a search model. This interpretation would be suitable for the other two versions of the model, which will only been hinted in the conclusive section of the Chapter.

[4] For instance, it can be thought that the two firms participate in an auction to 'buy' the market, thus the firm with higher productivity is the one that will offer the higher bid. However, this aspect will not be modelled.

case and getting good draws in the co-operative one, are generally not so large to alter this result.

A way to gain a further insight into the model is perhaps to illustrate its relationship with the structure of GH's original one: as mentioned above, in fact, the three stages of the present model can be matched with the three phases of GH's model. Then, the stage concerning the choice on the type of human capital in the present model corresponds to what is the *commitment* stage in GH's model; in fact, as will become clear later, the decision to invest in co-operative skills can indeed be seen as a commitment to a co-operative behaviour at the third stage, as by increasing the investment in co-operative skills at the first stage, the profitability of the competitive strategy at the third stage is decreased. The *random* stage in GH consists here of the determination of individual productivities; finally, what in GH is called the *action* stage is here interpreted as the decision on either setting up the joint enterprise or participating in the contest for the market, which is in turn affected by the commitments taken at previous stages and by the random determinations of individual productivities.

7.1.2 The 'Race for market demand'

I shall now proceed to the more detailed illustration of the model. Its timing is illustrated in Figure 7.1, where a brief description of the main actions involved, along with the indication of the relevant variables, be they under agents' or Nature's control, accompanies each stage of the game.

I believe that the appropriate way to illustrate the structure of the game is to start from the end, i.e. from the last two stages, and then proceed backwards. The basic choice agents have to make in stage 3 is whether to set up a joint enterprise or to compete against each other for success in the 'race for the market'. The first choice, which of course requires *both* agents to agree, implies that they share a cost function and split the profit in, say, equal parts. Therefore, the related payoff is for the *i*-player is:

$$\pi_i^C(Y;\theta_i,\theta_j)=\frac{1}{2}\{pY-c^C(Y;\theta_i,\theta_j)\} \qquad (7.1)$$

The superscript C refers to the 'Co-operative' enterprise, i.e. the one associated with joint production. Y is the quantity produced jointly by the two firms. The cost function $c^C(Y;\theta_i,\theta_j)$ depends on the two parameters θ_i and θ_j, which, as explained above, are the two individual productivities associated with the joint production function; they are determined by the move of Nature at the second stage and by agents' decisions on skill investment at the first stage. Total profits are here shared equally amongst the agents. Notice that the demand function facing the co-operative firm is assumed, for simplicity, to be infinitely elastic with respect to the price[5].

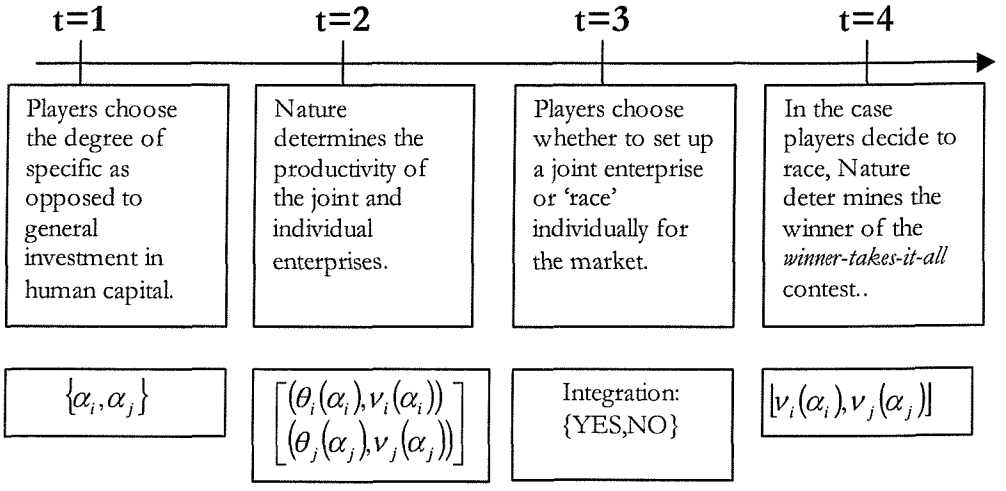

Figure 7.1

The other option the two players have is to enter the race for appropriating the market demand. In this case, individual payoff is:

$$\pi_i^R(y_i;v_i) = P_{WR}\{py_i - c_i^R(y_i;v_i)\} \qquad (7.2)$$

Here the superscript R denotes variables relative to the other option agents have, i.e. Racing for the market. P_{WR} represents the probability of winning the 'race', according to the rules that have been specified above. Given the winner-takes-all nature of the contest, the payoff associated with losing the race amounts to zero. Profits in case the race is won are equal to revenues, where the demand is the same as that for the co-

[5] This assumption is mainly made abstract away from the complications that would arise if firms were assumed to have some degrees of market power. In any case, though, since the co-operative firm and

operative case, net of costs $c_i^R(y_i;\nu_i)$. Costs are now only affected by y_i, which is the quantity produced by the individual firm that has won the contest, and individual productivity in the competitive case, i.e. ν_i. The interaction in the last stages is such that it suffices that one firm refuses to Co-operate that the contest takes place. As we shall say, this strategic aspect will be crucial to the results of the game.

In stage 2 the move of Nature has determined the parameters (θ_l , ν_l), $l=\{i,j\}$, with which agents approach the third stage. As already mentioned, they are the outcome of a random process, which is in turn affected by the type of investment in human capital made at the first stage. In order to make this relation clear, I will sometimes write the parameters as follows: $(\theta_l(\alpha_l),\nu_l(\alpha_l))$. A key point is that these are private information of each agent at the third stage, although each agent can attach a probability distribution to the parameters based on the action observed in the first stage. I will leave to a further discussion the case in which the agent cannot *infer* a probability distribution because of a situation of radical uncertainty on some other characteristics of the other party.

As already shown, the parameters θ and ν affect the cost functions, and thus determine the productivity of the joint and individual production functions respectively. In what follows, costs are given a simple quadratic form:

$$c^J(Y;\theta_i,\theta_j)=Y^2(1-g(\theta_i,\theta_j)) \tag{7.3}$$

$$c_l^R(y_l,\nu_l)=(y_l)^2(1-\gamma\nu_l),\ l=\{i,j\} \tag{7.4}$$

Equation (7.3) represents the cost function associated with the joint form of enterprise. It is made up of a standard quadratic term, and by a second term capturing the impact of the degree of complementarity among individual skills on joint productivity. More precisely, $g(\theta_i,\theta_j)$ is a function satisfying the following basic features:

$$g(\theta_i,\theta_j)\in[0,1)$$

$$\frac{\partial g(\theta_i,\theta_j)}{\partial\theta_l}\geq 0 \tag{7.5}$$

---

that winning the race would have the same degree of market power, the final results would not be

That is, $g(\theta_i, \theta_j)$ represent a bonus in the costs, which is non-decreasing in the size of individual co-operative productivity. In the remainder of the Chapter it will be called the *co-operative bonus*. The extent to which costs can be reduced depends on the way individual contributions to joint production are combined, i.e. on their degree of complemantiriety/substitutability. As anticipated, I will only consider the two extreme cases of perfect complements and perfect substitutes in individual contributions, which are represented respectively by the following specifications:

$$g(\theta_i, \theta_j) = \beta(\min\{\theta_i, \theta_j\}) \tag{7.6a}$$

$$g(\theta_i, \theta_j) = \beta(\max\{\theta_i, \theta_j\}) \tag{7.6b}$$

Equation (7.6a) describes a situation where individual skill investments are perfectly *complementary* in that both agents needs to realise a high value in their own θ_l in order to reap the bonus. That is, it suffices that one of the two agents gets θ_l=0 to eliminate the bonus. Expression (7.6b), instead, represents a situation in which investment in individual skills are perfectly *substitutes*, as it is sufficient that *any* agent obtains a high realisation in her own θ_l to permit the realisation of the bonus. β is instead a *technological* parameter that measures the size of the bonus, provided that a 'good' match has emerged from agents' actions and Nature's move. Condition (7.5) and the fact that θ_l can assume values of either zero or one imply that $\beta \in (0,1)$.

The *individual* cost functions (7.4) have a similar shape, in that they are made up of a 'standard' quadratic term, which of course is now a function solely of the individual quantity that is being produced, and by a second term that determines the *bonus* for individual costs when the parameter v_l has reached a high value. γ here plays the same role as β in (7.6) as it represents the size of the competitive bonus when a high realisation of v_l makes this possible. To be sure, given the 'winner-takes-all' nature of the race, the bonus can be appropriated only if the agent wins the race. Furthermore, I assume that the winner of the competition for the market is determined by the pair of parameters vs. More precisely, the winner is the individual who is endowed with the higher v, and the possible case of a tie is resolved with the

---

affected by this simplifying assumption.

toss of a (fair) coin. Therefore, an agent is sure to win (to lose) the race if he has got a high (low) v and her opponent has obtained a low (high) v. In all of the other cases, when agents draw the same v, they have probability equal to ½ of winning the race.

7.1.3 The Choice of Skills

I now turn my attention to the initial stages of the game. The first stage concerns the agents' choice on their degree of investment in *co-operative*, as opposed to *competitive*, skill. This distinction is akin that of Aoki (2000) between *malleable* and *functional* skills, on which he bases his account of the institutional differences between the Japanese and the US types of corporate organisation. This account is based on the intuition that different types of competencies need to be developed depending on the productive setting in which an agent operates. One may also think of co-operative and competitive skills as measuring the degree to which an agent has invested in searching for her suitable counterpart within a population of entrepreneurs. Although it may be generally thought that there are some complementarities between the two types of investment, I shall assume that there exists a sharp trade-off between the two, so that investing one additional unit of effort in one type of skill means that the same unit must be detracted from investment in the other skill. The best way to model these considerations is to assume that each agent has available a fixed, but divisible amount of resources, e.g. time, or money to be spent on education, which can be allocated to the development of the two skills.

Accordingly, the variable $\alpha_i \in [0,1]$ will measure the amount of *effort* put in the investment in co-operative skill, whereas $(1-\alpha_i)$ represents the amount of effort available to be invested in competitive skill. I assume that investing in skills bring about a cost for the agent, which will be represented by the function $d(\alpha_i)$. The shape of this function is not obvious, as it would depend on the sum of the costs associated with investing in the two skills. For now, I assume that the cost for competitive skill is always zero, so that only the other form of investment carries a cost for the agent. Therefore, skill investment cost is bound to be non-decreasing in α_i and to be zero when α_i is zero:

$$d'(\alpha_i) \geq 0, d(0) = 0 \qquad (7.7)$$

The choice of α_i has an impact on the determination of the parameters (θ_i , ν_i). In particular, I assume that the higher the effort put in the investment in a skill, the higher the probability that a high productivity is realised. In particular, the parameters are assumed to be determined by a simple binomial distribution:

$$\tilde{\theta}_i = \begin{cases} 1 & \text{with prob. } \rho\alpha_1 \\ 0 & \text{with prob. } (1 - \rho\alpha_1) \end{cases} \tag{7.8}$$

The parameter $\rho \in (0,1)$ is an index of the *uncertainty* present in the economy. In fact, it prevents the agent from being *absolutely* certain to obtain a high contribution to joint productivity even when devoting her whole effort to co-operative skill, i.e. when α_i=1. This parameter may be thought of as an effect of the latent uncertainty present in the economy, which may prevent the realisation of a good 'match' between agents because of their skills being incompatible.

The same argument applies to the other parameter affecting the productivity of individual production:

$$\tilde{\nu}_i = \begin{cases} 1 & \text{with prob. } \tau(1-\alpha_1) \\ 0 & \text{with prob. } 1 - \tau(1-\alpha_1) \end{cases} \tag{7.9}$$

Here, the parameter τ plays the same role as ρ in expression (6.8), and, again, effort put into competitive skill, i.e. $(1- \alpha_i)$, increases the probability of 'drawing' a high productivity.

For expositional purposes, I shall adopt the following terminology, which permits the unification of the two cases of investment in co-operative and competitive skills in a single framework. First, I shall refer to parameters θ and ν as *co-operative* and *competitive productivity* respectively, by analogy to the type of skill they are associated with. The same applies to the *bonuses* that may obtain in the two cases. Second, I shall say that when agents get a high level productivity in either of the cases, their investment in the related skill has been *successful.* This implies that under the specification (7.6.a) of the co-operative bonus, the investment of both agents has to be successful, whereas under (7.6.b) it suffices that one agent realises a successful investment in co-operative skill.

The main idea I wish to capture with this model is straightforward: the higher the investment in co-operative skill, the higher the probability of realising a good match in

the case of setting up a joint enterprise, but the lower the probability of selecting a high productivity for the individual case and of winning the race. Conversely, the higher the effort devoted to the development of competitive skill, the higher the probability of winning the race, but the lower that of carrying out a good match in the joint enterprise. For instance, devoting all effort to co-operative skill, i.e. $\alpha_i=1$, makes as high as possible the probability that individual contribution to joint production is high, but causes v_i to be zero. In fact, this does not determine that the race would certainly be lost, as the agent may win the toss of the coin if confronted by an opponent who also had a zero competitive productivity.

Therefore, the *ex-post* probability of winning the contest varies in relation with the draw of v:

$$P_i(WR|v_i=1)=P(v_j=0)+\frac{1}{2}P(v_j=1)=1-\frac{1}{2}\tau(1-\alpha_j) \tag{7.10}$$

$$P_i(WR|v_i=0)=\frac{1}{2}P(v_j=0)=\frac{1}{2}(1-\tau(1-\alpha_j)) \tag{7.11}$$

Hence, the *ex-ante* probability of winning the contest, which depends on α_i is:

$$P_i(WR)=\frac{1}{2}(1+\tau(\alpha_j-\alpha_i)) \tag{7.12}$$

Notice that, whenever α_i is different from α_j, a decrease in τ implies an increase in the variance of winning the race, which confirms the interpretation given above of τ as an index of the uncertainty present in the economy, in particular with respect to the individual production functions.

7.2 THE SOLUTIONS TO THE GAME

7.2.1 Optimal Strategies in the Subgame with Complementary Efforts in Joint Production

First, I deal with the case of complementary efforts within the joint production function (eq. 7.6.a), which requires *both* agents to draw a high co-operative productivity. I leave the case of substitute efforts to the next section. Consider the strategic situation from the third stage of the game: here each agent is aware of the draw of her pair of parameters, but is uncertain as to that of her counterpart. Though,

the agent can build a probability distribution on this, which is based on the observation of the action carried out by her counterpart in the first stage. Hence, she will form a belief on her partner's pair of productivity parameters that is based on (7.8) and (7.9). Now, the decision agent i has to make concerns whether to offer her availability to set up a joint enterprise or to go directly for the race. I shall refer to these actions as Co-operating and Racing, respectively, with labels C and R, even though whether the joint enterprise will be constituted depends on the co-operation of *both* agents. In other words, action C only signals the *availability* of an agent to set up the joint enterprise. In the case the other agent refuses to Co-operate, she will be called to race. In fact, recall that the availability of *both* agents is required in order to form the joint enterprise.

Agent i's expected payoff from Co-operating in the joint enterprise varies with hers and her opponent's draws of θ_l. In what follows I show the general solution to the problem of maximising profits. The optimal quantity is:

$$Y^* = \frac{p}{2\left(1 - g\left(\theta_i, \theta_j\right)\right)} \tag{7.13}$$

Hence, according to (7.1), profits assigned to each player are:

$$\pi_l^C\left(\theta_i, \theta_j\right) = \frac{p^2}{8\left(1 - g\left(\theta_i, \theta_j\right)\right)} \quad l = i,j \tag{7.14}$$

By solving the same problem in the case of competition and substituting in (7.2), one obtains that the payoff available for the firm that wins the race is given by:

$$\pi_l^R\left(\nu_l\right) = \frac{p^2}{4\left(1 - \gamma\nu_l\right)} \quad l = i,j \tag{7.15}$$

Thereby, this is the 'stake' in the race for the market. Notice that, in this case of complementary contributions within the co-operative production function, neither before nor after observing her own productivity can an agent be sure about the relative magnitude of $\pi_l^C\left(\theta_i, \theta_j\right)$ and $\pi_l^R\left(\nu_l\right)$. In fact, the former depends on the realised value of θ of her opponent, which is unknown until both agents decide to Co-

operate[6]. Moreover, it is clear that the selection of high productivity in both co-operative and competitive skills has the effect of increasing the relative expected payoff, and thus the attractiveness, of the related action.

Since there are four possible outcomes in individual productivities, we have to consider each different case in turn.

CASE (1): $(\theta_i=1, \nu_i=1)$

In this case the agent has high productivity in both joint and individual production. Therefore, her expected payoff if both agents Co-operate is:

$$U_i(C,C;\theta_i=1)=\rho\alpha_j\pi_i^C(\theta_i=1,\theta_j=1)+(1-\rho\alpha_j)\pi_i^J(\theta_i=1,\theta_j=0)= \\ =\left(\frac{p^2}{8}\right)\left(\frac{1-\beta(1-\rho\alpha_j)}{1-\beta}\right) \tag{7.16}$$

Expected payoff from competing is given by:

$$U_i(R,\cdot;\nu_i=1)=\frac{\tau(1-\alpha_j)}{2}\pi_i^R(\nu_i=1)+(1-\tau(1-\alpha_j))\pi_i^R(\nu_i=1)= \\ =\left(\frac{p^2}{8}\right)\left(\frac{2-\tau(1-\alpha_j)}{1-\gamma}\right) \tag{7.17}$$

Therefore, it will be profitable to select the first option if and only if:

$$U_i(C,C;\theta_i=1)\geq U_i(R,\cdot;\nu_i=1)\Leftrightarrow\alpha_j\geq\bar{\alpha}_1 \tag{7.18}$$

where

$$\bar{\alpha}_1=\frac{(1-\beta)(1+\gamma-\tau)}{\rho\beta(1-\gamma)-\tau(1-\beta)} \tag{7.19}$$

The intuition is quite clear: a player Co-operates if and only if she observes a high enough investment in co-operative skill in her opponent at the first stage. In fact, this ensures that the co-operative bonus in joint productivity is sufficiently likely, thus making joint production preferable to individual production. In other words, if the observed effort in co-operative skills that the other agent has undertaken is sufficiently high, then the expected payoff from Co-operating exceeds that of Racing. This holds, at least for sufficiently high values of β[7], notwithstanding the fact that the agent has

---

[6] In fact, the inequality that determined whether the payoff at stake in Racing exceeds that obtainable when Competing is: $\gamma\nu_i>2g(\theta_i,\theta_j)-1$.

[7] In fact, (7.18) holds under the condition that the denominator is positive, which I shall assume throughout the analysis. This is the case if β outstrips a threshold level that depends on γ; more

also obtained high productivity in individual production. The threshold level depends on the array of parameters of the model, and it is clear how an increase in β has the effect of decreasing such a threshold, as the expected payoff from co-operative production increases, whereas an increase in γ, for the same reasons, has an opposite effect on $\bar{\alpha}_1$. Moreover, higher uncertainty in the realisation of θ - namely, an increase in ρ - has a negative effect on the threshold, whereas greater uncertainty in the occurrence of γ, i.e. an increase in τ, has the opposite effect on $\bar{\alpha}_1$.

CASE (2): *(θ_i=0, ν_i=1)*

If the agent has got a bad draw in co-operative productivity, that is θ_i=0, she is then certain to obtain the following payoff, no matter what her opponent's draw is:

$$U_i(C,C;\theta_i=0)=\left(\frac{p^2}{8}\right) \tag{7.20}$$

Comparing this with expression (7.17), one obtains that $U_i(J,J;\theta_i=0)<U_i(R,\cdot;\nu_i=1)$ for any value of α_i. Therefore, if the agent gets a good draw in competitive productivity and a bad one in co-operative productivity, she will always opt to Race. In fact, under this situation the *i*-player is certain *not* to gain the bonus in co-operative production, whereas she has positive probability of winning the Race and the competitive bonus. This makes Racing the best option in this case.

CASE (3): *(θ_i=1, ν_i=0).*

Conversely, if the *i*-player gets high productivity in joint production and low in individual production, then she will always opt to Co-operate. In fact, her expected payoff from competing for the market is now:

$$U_i(R,\cdot;\nu_i=0)=\left(\frac{p^2}{8}\right)\left(1-\tau\left(1-\alpha_j\right)\right) \tag{7.21}$$

Hence, comparing (7.21) with (7.16) shows that choosing not to Race is always optimal, no matter what the opponent's individual choice of α_j is. The intuition is analogous to case (2) above.

---

precisely, the condition is $\beta>\frac{\tau}{\tau+\rho(1-\gamma)}$. If this condition did not hold, then the agent would always select 'Race' in this case and no investment in co-operative skill at the first stage, thus making the

CASE (4): $(\theta_i=0, \nu_i=0)$.

Finally, comparing (7.21) with (7.20) demonstrates that Co-operating is always optimal for all positive values of α_j, and at least as optimal when α_j is equal to zero. Therefore, the agent will choose to Co-operate when the agent gets bad draws in both productivities.

On the grounds of this analysis, we can summarise as follows the agent's optimal strategy, which depends on the productivity draw and on the observed effort in co-operative skills by her counterpart :

$$s_i^*(\theta_i(\alpha_i),\nu_i(\alpha_i);\alpha_j)=\begin{cases}\begin{cases}C & \text{if } \alpha_j \geq \overline{\alpha}_1\\ R & \text{otherwise}\end{cases} & \text{if } \theta_i(\alpha_i)=1,\nu_i(\alpha_i)=1\\ C & \text{if } \theta_i(\alpha_i)=1,\nu_i(\alpha_i)=0\\ R & \text{if } \theta_i(\alpha_i)=0,\nu_i(\alpha_i)=1\\ C & \text{if } \theta_i(\alpha_i)=0,\nu_i(\alpha_i)=0\end{cases} \qquad (7.22)$$

In summary, whenever the agent gets a bad draw in competitive productivity she will always opt to Co-operate (or, to be more precise, she will 'offer' to Co-operate: remember that the selection of the co-operative enterprise always requires both agents to offer their availability to Co-operate). However, when she gets high competitive productivity, she will opt for Racing when her co-operative productivity is low, and she will condition her choice on the observed investment in co-operative skills of her counterpart when both her productivities are high. In fact, when competitive productivity is low, the perspective of losing the Race and earning a payoff of zero with high probability makes this a rather unattractive option: in this case it is better for the agent to go for the 'safer' option of Co-operating, which ensures a positive payoff even when no co-operative bonus is delivered.

When competitive productivity is high, instead, the chances to win the race with the relative bonus are quite high, thus making this quite an attractive option. This will in fact be the agent's option when her co-operative productivity is low, which ensures that no co-operative bonus will be delivered. However, when her co-operative productivity is high, there is a positive probability of earning the bonus: this will

---

model quite uninteresting.

depend on her counterpart's co-operative productivity, which the agent cannot observe but which can estimate on the grounds of his investment in co-operative skills carried out at the first stage. Hence, if she has observed high enough effort by her counterpart, the probability of earning the co-operative bonus gets sufficiently high, thus making this the most preferred option, at least when β is sufficiently high (see note 5). This analysis is consistent with the intuition given at the outset (section 7.1.1) on the higher risk associated with competing for the market with respect to co-operating.

To be sure, the presence of unconditional choices in three out of four cases depends on the particular cost functions that have been adopted and on the simplicity of the binomial distribution scheme for determining productivity. However, I hope that this simplification helps to focus on the main aspects of the relationship between risk and production that were laid down at the outset.

7.2.2 Optimal Strategies in the Subgame with Substitute Productivity in Joint Production

This case differs from the preceding one only for the different shapes of the joint production bonus, which is now expressed by (7.6.b). This represents a situation in which individual contributions to joint production are *substitutes*, as it suffices that only one agent's investment in co-operative skills is successful in order to yield the co-operative bonus. Although the overall 'stake' of joint production remains unchanged as in (7.14), this situation is different from the strategic point of view. This can be shown by noticing the changes in the payoffs functions in the joint case: if the agent gets a good draw in joint productivity, she is now certain to earn the highest payoff available:

$$U_i(C,C;\theta_i=1)=\left(\frac{p^2}{8}\right)\left(\frac{1}{1-\beta}\right) \tag{7.23}$$

Conversely, when her investment in co-operative skill is not successful, she may still hope to earn the bonus if her partner's investment turns out to be successful:

$$U_i(C,C;\theta_i=0)=\left(\frac{p^2}{8}\right)\left(\frac{1-\beta(1-\rho\alpha_j)}{1-\beta}\right) \tag{7.24}$$

Notice that this expression is equal to (7.16), which was associated with a good, rather than a bad, draw in joint productivity. If the costs that the agent has to sustain to engender co-operative skill are sufficiently high, then, a typical free-rider problem arises, as each agent has the incentive to free-ride on other's investment in co-operative skill. On the other hand, an agent is more likely to Co-operate at the third stage if she is hopeful that her counterpart's investment in co-operative skill has been successful.

Some simple comparisons similar to that carried out in the previous section leads to find out the third stage optimal strategy. In general, its exact form depends on the parameter values, but the strategically more interesting version[8] is as follows:

$$s_i^*(\theta_i(\alpha_i),\nu_i(\alpha_i))=\begin{cases}\begin{cases}C & \text{if } \alpha_j \le \bar{\alpha}_2 \\ R & \text{otherwise}\end{cases} & \text{if } \theta_i(\alpha_i)=1,\nu_i(\alpha_i)=1 \\ C & \text{if } \theta_i(\alpha_i)=1,\nu_i(\alpha_i)=0 \\ \begin{cases}C & \text{if } \alpha_j \ge \bar{\alpha}_3 \\ R & \text{otherwise}\end{cases} & \text{if } \theta_i(\alpha_i)=0,\nu_i(\alpha_i)=1 \\ C & \text{if } \theta_i(\alpha_i)=0,\nu_i(\alpha_i)=0\end{cases} \tag{7.25}$$

where:

$$\bar{\alpha}_2=\frac{2\beta+\tau(1-\beta)-(1+\gamma)}{\tau(1-\beta)} \tag{7.26}$$

$$\bar{\alpha}_3=\frac{(1-\beta)(1+\gamma-\tau)}{\rho\beta(1-\gamma)-\tau(1-\beta)} \tag{7.27}$$

In this case, when the agent gets a low competitive productivity, i.e. $\nu_i=0$, then she will always opt to Co-operate as well as in the previous case, thus the intuition is the same as before (see section 7.2.1). However, when competitive productivity is high, i.e. $\nu_i=1$, then the situation changes quite considerably. In fact, when the agent gets good draws in both productivities, i.e. $\theta_i=1$and $\nu_i=1$, then she is certain of earning the co-operative bonus independently from her opponent's action. However, at least when γ is sufficiently high[9], the option of competing for the market and earning the relative

---

[8] See next note to see in which sense this is true.

[9] In fact, the dichotomic choices that are indicated in (7.25) are subject to some constraints in the parameters. Thus, when θ_i and ν_i are both equal to 1, then the agent will always opt for Co-operating

bonus will be attractive as well. The chances of winning the race depend on the *j*-player's investment in competitive skill: if this is relatively low, i.e. α_j is high, the *i*-player will expect with high probability to win the race, thus, at least for some ranges in the parameter values, she will opt to Race. This explains why the sign of the inequality in the first possible case of (7.25) is reversed with respect to (7.22): the *i*-player will now decide to Co-operate only if the other agent has performed a sufficiently *low* amount of effort: the probability of winning the race must not be too high in order to make Co-operation attractive.

Furthermore, when the agent's productivity is high only in competitive production, the agent's final choice is now conditional on her counterpart's action at the first stage. Notice, instead, that this situation would have unambiguously determined Racing in the previous case. In fact, if the counterpart's effort in co-operative skill is sufficiently high, then the probability of gaining the bonus in the case of Co-operation could be sufficiently high to make Co-operation the best option, notwithstanding that the chances of winning the Race and collecting the relative bonus are high too. Instead, if α_j is below the relative threshold, then the probability of gaining the bonus are relatively low, thus the agent will decide to Race.

7.2.3 The Ex-Ante Choice

Depending on the players' draws of productivities, we obtain 16 possible outcomes in the subgame starting from the third stage of the game, which correspond to all the possible combinations that can occur between the realisations of the pairs of productivities for the two agents. The associated optimal individual actions are those indicated by (7.22) and (7.25), respectively, for the two cases analysed of complementary and substitute inputs in joint production. Here I assume that agents have a sufficient level of rationality enabling them to solve the game through backward induction. Therefore, they are able to work out their opponent's optimal decisions from the third stage, to compute the probability of each possible outcome

---

if $\beta \geq \frac{1+\gamma}{2}$, and apply the rule indicated in (7.25) for all other values. Similarly, if $\beta \leq \frac{\tau}{\tau + \rho(1-\gamma)}$ the agent will always Race when v_i is 1 and θ_i is 0, while following (7.25) in all other cases. The rule depicted in (7.25), thus, represents the more interesting case from the strategic point of view.

depending on the observed fist stage action, and then to compute the expected payoffs derived from each first stage option. A compact expression for the objective function to be maximized *ex-ante* by an *i*-player is given by the following expression:

$$U_i(\alpha_i,\alpha_j;S_i^*(\alpha_i,\alpha_j),S_j^*(\alpha_i,\alpha_j))=\tilde{\pi}_i(s_i^*(\alpha_i,\alpha_j),s_j^*(\alpha_i,\alpha_j))-d_i(\alpha_i) \quad (7.28)$$

$s_i^*(\alpha_i,\alpha_j)$ is the ex-post optimal strategies for both players, that is given by (7.22) or (7.25) depending from the case we are dealing with. Notice that both the agent's own optimal choice and her opponent's one are incorporated into the ex-ante payoof function. $\tilde{\pi}_i$ represents the expected ex-ante payoff from selecting a certain action α_i at the first stage, *given* α_j:

$$\tilde{\pi}_i(s_i^*(\alpha_i,\alpha_j),s_j^*(\alpha_i,\alpha_j))=\sum_{\theta_i=0}^{1}\sum_{\upsilon_i=0}^{1}\sum_{\theta_j=0}^{1}\sum_{\upsilon_j=0}^{1}\pi_i^*(s_i^*(\theta_i(\alpha_i),\nu_i(\alpha_i)),s_j^*(\theta_j(\alpha_j),\nu_j(\alpha_j))) \quad (7.29)$$

Notice that in (7.29) all the sixteen possible outcomes are taken into account by agent *i*, and that π_i^* represents the (expected) payoff associated with agent *i*'s optimal action, as computed in expressions (7.16), (7.17), (7.20), (7.21), (7.23), (7.24). Finally, the payoff in (7.28) is also affected by the cost function of investment in co-operative skills, which is consistent with the discussion set out in section 7.1.3 and condition (7.7).

7.2.4 The Equilibria of the Game

Rather than presenting the *closed-form* solution to the game, which would get us entangled in tedious computations and countless limitations on parameters, I choose to present the results of *numerical* analysis I have conducted; in particular, I will present what look like generic patterns of equilibria that persist across significant changes in the parameters, which will be grouped into different scenarios. In fact, despite the rather complex structure of the model, the final solutions can be sorted in very simple types of interactions. Moreover, in the present section I will further simplify the model by assuming that there are only two levels for effort α that are feasible, namely 0 and 1. This permits me to present the normal-form interaction, and the corresponding equilibria, as a two-person game with two available strategies. Some examples of the more general case where α is a continuous variable on the interval [0,1] could be easily

constructed; however, the numerical investigation clearly shows that the basic insights of the interaction can be captured under this simplifying hypothesis.

7.2.4.A Prisoner's Dilemma Scenario

The first scenario that can be observed is that of a Prisoner's Dilemma. It is worth noticing that this obtains both under the hypotheses of complementary and substitute efforts. Under the former, the following normal-form game obtains after having solved the game through backward induction, and computing the expected payoffs associated with the pairs of effort levels available to the players. The key parameters for the following analysis are the co-operative and competitive bonuses, which in the present case are γ=0.4 and β=0.7 [10]. That is, the joint technology is more efficient than the individual one in that it allows a greater bonus, and thus a greater total amount of profits to be shared and a higher quantity produced[11].

	α_j=1	α_j=0
α_i=1	27, 27	-2, 36
α_i=0	36, -2	21,21

Figure 7.2

The reason why a Prisoner Dilemma's type of situation occurs is clear if we think of the implications of the complementary effort hypothesis in (7.6.a). First, it is apparent that if the counterpart does not invest in co-operative skills, than the agent can be certain that the co-operative bonus will not obtain, thus she has no incentive to invest either. Second, whether it is appropriate to invest in co-operative skill when the other

---

[10] In particular, in the complementary case, the payoffs in Figure 7.2. obtain under the following set of parameter values: {p=10 ; γ = .4 ; β=.7 ; τ=0.8 ; ρ=0.8; δ=2}. However, it can be shown that this, like any other scenario, obtains for a large set of the parameters values, thus the results here described are generic.

[11] This, of course, is true in an 'ex-post' perspective, which compares the two technologies supposing that the productivities realisation has been the one permitting the occurrence of the bonus. The payoffs reported in the matrix of the game, instead, reflect the ex-ante expected payoffs. Under this

agent does, depends on the stakes in play when Co-operating and Racing. If the first is not large enough, than the incentive to Race, and thus be likely to win the contest and the entire market, outweighs the payoff from Co-operating, where the payoffs, even though augmented by the bonus, will be shared between the two agents. In fact, given that the other agent has not invested in competitive skills, i.e. α_j is equal to 1, the *i*-player is relatively likely to win the race for the market and get the bonus if she, on the contrary, does invest in competitive skills, i.e. she sets α_i equal to 0. Therefore, for meaningful values of β and γ, the incentive to take advantage of the counterpart's low investment in competitive skills is too high, and the agent will prefer to Race. In the game represented in Figure 7.2, therefore, the opportunity to win the race when the other agent is certain to draw a low competitive productivity outstrips the benefits of Co-operation. The fact that *joint* investment in co-operative skills would instead boost individual expected payoff above the payoffs attained in the case of investment in competitive skills, makes this situation exactly alike a Prisoner's Dilemma: the incentive to profit from the counterpart's decision of investing in co-operative skills causes the final outcome to be a typical no-co-operation trap.

A Prisoner's Dilemma can also arise when joint efforts are substitute, as in (7.6.b). The intuition is quite similar to that just illustrated: the incentive not to invest in co-operative skills when the other agent is doing so, thus considerably increasing the probability of winning the race, may outstrip the bonuses of Co-operation, even if in this case the probability of obtaining the co-operative bonus is higher than in the previous case as a single successful investment in co-operative skills is sufficient. On the other hand, if the other agent does not invest in co-operative skills, then investing would expose the agent the similar risk of being dragged to the race with very low probability of winning it, thus the optimal strategy is not to invest.

7.2.4.B Symmetrical Co-ordination Game Scenario

It suffices to slightly increase the extent of the co-operative bonus with respect to the previous case to make the structure of the interaction completely different. In fact,

---

perspective, the efficiency must be measured with respect to the action available at the first stage, i.e. in terms of the choice on the level of investment in co-operative skill.

leaving γ equal to 0.4 and all the other parameters as in note 9, but increasing β to 0.8, brings about a change in the 'structure' of the interaction. The expected payoffs, under the perfect complement case, are in fact as in Figure 7.3:

	$\alpha_j=1$	$\alpha_j=0$
$\alpha_i=1$	40,40	-2, 36
$\alpha_i=0$	36, -2	21,21

Figure 7.3

The change in β causes the co-operative bonus to be relatively large, so that investing in co-operative skills now becomes the better option when the counterpart does so. However, for the same reasons highlighted before, the agent has no incentive to invest if the other agent does not invest. This makes the structure of the interaction like a co-ordination game, where even the sub-optimal outcome when no agent invests in co-operative skills is indeed an equilibrium of the game. Notice that this type of co-ordination game only occurs for complementary efforts in joint production.

7.2.4.C Hawk-Dove Scenario

When efforts are complementary in joint production, in fact, a different type of co-ordination problem arises, as shown by Figure 7.4[12]:

	$\alpha_j=1$	$\alpha_j=0$
$\alpha_i=1$	32, 32	27, 36
$\alpha_i=0$	36, 27	18, 18

Figure 7.4

Here, the co-operative bonus is sufficiently high to make the agent willing to invest in co-operative skill even when the other agent is not doing so. Moreover, such an agent is sure that her counterpart will not Race at the third stage, as the very observation of

the investment in co-operative skill makes Co-operating the optimal strategy for the counterpart's agent at stage 3 (recall the optimal strategy scheme in (7.25)). Therefore, the agent who invests in co-operative skill is sure that her investment will pay off in terms of higher probability of getting the co-operative bonus. However, this situation causes an asymmetry in the payoffs, as obviously the agent who invests in co-operative skills has to sustain the relative costs, whereas the other agent simply free-rides on the other player's investment.

7.2.4.D Dominant Strategies Scenarios

As mentioned above, I have been focussing on the cases that I thought were more interesting from the strategic point of view, i.e. those in which a player's action depends on the observation of the degree of commitment of the other player on the co-operative skills, or in which inefficient situations like the Prisoner's Dilemma occurs. However, for the completeness of the analysis, I have to mention that when the difference between the two relevant parameters that I have identified, i.e. β and γ, becomes relatively large, then the structure of the best reply functions (7.22) and (7.25) modifies in such a way as to make one action *unconditionally* the optimal one, *and* the related outcome the efficient one. In this sense, thereby, these cases differ from the first situation that has been illustrated, i.e. the Prisoner's Dilemma, as not only does individual behaviour follow dominant strategies, but also the corresponding result is efficient. Hence, if ν is large enough with respect to β, and the corresponding level of uncertainty τ is comparable to ρ, then Racing will become the more preferred option and also the one guaranteeing the Pareto-superior outcome. The opposite occurs when, instead, it is β to outstrip ν by a large amount.

---

[12] The associated parameter values are: {p=10 ; γ = .3 ; β=.7 ; τ=0.8 ; ρ=0.8; δ=3}

7.3 The Relationship Between Social Norms and Technology

7.3.1 Some Introductory Considerations

The analysis of the foregoing section has shown that the structure of the interaction, once the game has been solved through backward induction and is reduced to its normal form, can be brought down to some basic types of interaction, such as the Prisoner's Dilemma, and a symmetric and asymmetric co-ordination game. This makes the investigation of the impact of social norms on the economic outcome particularly straightforward, in the light of the analysis developed in the third and fourth chapter of the thesis. In fact, I will stick with that model in assuming that, on the one hand, social norms play a part in individual motivations, in that an other-regarding inclination to reciprocate others' conformity to a moral standard, however understood, is a basic motive to action entrenched in the individual system of ends. On the other hand, norms are not permanent, but they are the evolutionary outcome of repeated strategic interaction between (boundedly) rational individuals who weigh up self-interested and (conditional) other-regarding motivations within their system of ends.

In what follows, then, I shall simply interpret the present model in the light of the foregoing analysis, thus hoping to offer a novel perspective on the subject of the relation between norms and economic growth. Before starting off, though, I need to put forward two methodological caveats. First, in what follows I will use an evolutionary argument in accounting for the relationship between norms and growth. Although a formal analysis will not be provided with respect to this point, it is evident that this would be a straightforward extension of the present model. In fact, the two-person game analysed above could well be seen as the basic stage game of an evolutionary model. That is, at each instant of time a pair of agents belonging to two different populations would be randomly matched to play this game, as their choice would be for various reasons pre-determined. Hence, the static Nash equilibria of the basic game may be interpreted as the outcome of an evolutionary process in which agents slowly replicate what at each instant of time is the more advantageous action.

Moreover, one could suppose that all agents were matched at each instant of time, thus determining the *aggregate* outcome for the whole of the population. The macroeconomic evolutionary model developed in Chapter 4 and 5 could be seen as an example of such an argument, although the structure of that model does not conform with two-person interactions, as the interaction involves all the players together.

In this setting, the technological parameters associated with the two technologies affects the 'payoffs' of the game, thus determining the 'basins of attractions' of the different equilibria. For instance, an increase in the relative profitability of the co-operative technique with respect to the other will *enlarge* the basin of attraction of the related equilibrium, thus making it more likely that a co-operative social norm emerges as well. In other words, technological parameters and conditions of production act as economic incentives in shaping the evolutionary path of the social system, so that they have an impact on the type of social norms that will emerge.

The relationship between social norms and technology varies with the different scenarios we have found. In what follows, I analyse them in turn.

7.3.2 *Norms of co-operation in the Prisoner's Dilemma scenarios*

This is perhaps the case in which the impact of social norms is most effective: in fact, the analysis developed in chapter 4 can be immediately carried over to this case. This makes it possible to rely on the emergence of 'norms of co-operation' that now make the efficient outcome – namely, co-operation by both players in the joint production technology - a *possible* equilibrium, along with the self-interested-based one illustrated in section 7.2.4.A. In this setting, by *norm of co-operation* I mean a regularity of behaviour that makes *investing in co-operative skill* at the first stage of the game the strategy to which players of either population wish to conform. As this strategy is clearly contrary to agents' self-interest, it must stems from some type of other-regarding motivation, such as those I had illustrated in Chapter 4, i.e. normative expectations and/or mutual conformity to public interest. As suggested above, the relative size of the technological parameter will have an impact on the shape of the regions of attraction of the two equilibria, thus influencing the likelihood and the speed of convergence towards either outcome.

The intuition behind this result is straightforward: since this outcome is Pareto-superior, then every possible deviation from it by the agent would come to be judged as contrary to the expectations of other members of the community under the standard resentment hypothesis (section 4.2 and 4.3), and contrary to public interest under the model of conditional conformity to morality (section 4.4). In fact, in all of these cases, the choice not to Co-operate by an agent would inflict a material cost on her opponent, thus eliciting his resentment on the agent performing that action. Hence, the sense of resentment that the individual would experience in not living up to others' expectations, or in failing to conform to the general established norm, under the conditions that the agent is resentment-inclined – i.e., her propensity to other-regarding behaviour lies in the region given by conditions (3.14) and (3.17) – would suffice to offset the material loss incurred in renouncing to Race when it would be appropriate, in the self-interested sense, to do so.

On a more technical ground, there is another interesting point that can be drawn by applying the analysis I brought out in section 4.2 and 4.4 for the Prisoner's Dilemma. This concerns the relationship between the uncertainty moulding economic activities and the likelihood with which norms of co-operation emerge. One of the results of that chapter was that the type of equilibrium under the standard resentment hypothesis case (sec. 4.2) changes according to whether the highest incentive to defect was when the other party Co-operated or Defected. With what now may look like a slightly confusing definition, we called the first situation one of *substitute strategies*, whereas in the present chapter we have talked about substitute *efforts* to characterise the way individual contributions affect productivity in the joint technology. In fact, the former distinction refers to the whole strategic interaction underlying a Prisoner's Dilemma, whereas the latter only looks at the way individual contributions are combined in the joint production function to determine economies of scale and productivity. The point I would like to make is that, on the grounds of the numerical simulations that have been conducted, it seems that the only case that appears possible is that of *complementary strategies*, i.e. when not investing in co-operative skill and then going to Race is more appropriate when the other agent is not investing rather then

when he is co-operating. This was in fact the type of situation shown by the numerical example given in Figure 7.2.

The economic intuition for this result is that the Racing strategy is generally more *risky* than the Co-operative strategy. In fact, as noted in section 7.1.1, the 'winner-takes-all' nature of the contest associated with Racing, makes the relative outcome in the 'bad' state of the world much lower than the corresponding case in the co-operative outcome: in the latter case, in fact, the agent is always guaranteed a positive payoff even when the co-operative bonus is not realised. This explains why the situation appears to be one of complementary strategies: when the other agent does not Co-operate, the expected opportunity cost of Co-operating is higher than what expected when the other agent is instead Co-operating.

This has a straightforward implication in terms of the type of equilibria that obtain. We have found out that *anti-reciprocal* equilibria and *reciprocal equilibria* (see section 4.2) can only arise in conjunction with substitute strategies type of interactions: therefore, they can be ruled out as unfeasible in the present setting[13]. This result is suggestive that the 'paradoxical' outcomes of 'exploitation' of one class of agents by the other, typical of the anti-reciprocal equilibria, cannot occur. On the contrary, the only type of outcome that is now possible is the *reciprocal* one (see section 4.2 and Figure 4.5). These are characterised by that the amount of co-operation performed by the two populations is approximately the same. The intuition for this result hinges upon the considerations set out above concerning the distribution of risk between economic activities: since competitive equilibria are more risky, the 'material', or self-interested, incentive to deviate from the anti-reciprocal outcome for the agent who is giving in would be too high, so that outcome could not possibly be sustained as an equilibrium based on normative expectations or a concept of public interest. Conversely, since risk is lesser within co-operative productive activities, and deviation becomes less attractive in terms of material utility, then outcomes of mutual conformity to co-operation are more likely to be sustained as equilibria.

---

[13] This is true if the λs lie in the 'intermediate' region of values indicated in conditions (3.14) and (3.17).

This analysis shows clearly, although admittedly rather abstractly, the relevance of economic incentives associated with different technologies for the determination of whether co-operative social norms may emerge. It indicates that co-operative norms are more likely to emerge in association with less risky activities, i.e. when the material incentive to deviate is less significant.

Another insight can be obtained if one looks at the dynamical aspect of the model, by applying the evolutionary type of argument. In fact, if the economy starts off arbitrarily away from the co-operative equilibrium, though within its basin of attraction, one will observe that the frequency of adoption of co-operative norms and co-operative productive activities grows and reinforces each other along the convergence path. Thus, the process of convergence towards the equilibrium will be characterised by the joint diffusion of norms and economic activities of the same type. This aspect highlights the feedback of norms on technological adoption: given the propensity of agents to endorse public interest when other agents are doing so, and provided that the co-operative technology is identified by the agents as more efficient than the other, which is the case for a significant set of parameter values, then it will be their other-regarding prompt to action to elicit the adoption of the co-operative technology.

In this sense, then, it holds what argued earlier: norms affect technology adoption by fostering the adoption of technologies that have a 'compatible nature' with them, e.g. co-operative technology and co-operative social norms. Alternatively, competitive technologies and norms eliciting the use of competition among individuals in the determination of allocations, would show in the present model the same dynamical pattern of mutual 'self-enforcement'. Of course, the model has been built in such a way as to show this relationship quite clearly: in reality it is not so easy to associate the 'character' of a social norm with that of a technology. However, the present model suggests that there are several 'objective' parameters that can be analysed in order to shed light on this particular aspect, such as the degree of complementarity/substitutability of individual contributions within aggregate production function, the relative size of returns to scale in the co-operative and competitive production functions, the degree of uncertainty in the 'successful'

realisation of one's own human capital, in either type of skill. Hence, I am confident that further empirical content can be given to the present model.

7.3.3 Norms in Co-ordination Games Scenarios

Under both types of Co-ordination problem scenarios, i.e. symmetric and asymmetric (e.g. Hawk-Dove) situations, the consideration of other-regarding reasons to action does not change radically the structure of the interaction; neither does it make it possible to reach new types of equilibria, possibly Pareto-superior to standard ones. Therefore, one cannot expect the type of dramatic changes observed in relation with the Prisoner's Dilemma scenarios.

However, the concern for other-regarding motives to action may indeed have an impact on the interaction in reinforcing and speeding up the process of convergence toward one of the two available outcomes. In fact, it is well known (Weibull (1995: Ch. 1)) that in co-ordination problem the phase plan is divided into two regions associated with the basins of attraction of the two equilibria of the game. Since in this setting self-interested motivations go hand-in-hand with other-regarding ones, as this is a type of interaction of the 'mutually beneficial' kind (see sec. 2.1), then the presence of either normative expectations or public interest compliance will have the effect of reinforcing the cumulative, 'snowballing', process typical of evolutionary dynamics. To be sure, this characteristic has its pros and cons, as it clearly can decrease the time it takes to converge with respect to either the Pareto-superior or the inefficient outcome. Consequently, the event of lock-in to inefficient steady states that we observed in Chapter 4 and 5 could be speeded up, as an effect of other-regarding considerations. Likewise, these could also play a significant part within path-dependent stochastic processes such as those studied by Arthur (1985), in breaking the balance between two possible equilibria in favour of the inefficient one.

A final aspect to be dealt with concerns the role of asymmetries in co-ordination games. As shown by Weibulll (1995: Ch. 1), the overall type of evolutionary dynamics is different for symmetric and asymmetric co-ordination games: in the first case, the two equilibria of the stage game are attractors of the system even in a single-population setting. Conversely, in the second case, which corresponds to the scenario described in section 6.2.4.C, in a single population setting the only (stable) steady state

for system is the fixed strategy equilibrium of the stage game. This is clearly an inefficient solution. The only way to escape from this new form of lock-in requires the general recognition by the players of some asymmetries in the game, which makes it possible to 'label' them differently and thus it paves the way to feasible asymmetric steady states. This analysis is clearly reminiscent of Sugden's early contributions on the subject (Sugden (1986); see also Hargreaves-Heap and Varoufakis (2002) for experimental evidence on this subject).

In the present setting, the shift from a single to a two-population setting would require one of the two groups of agents to give in systematically to the other. Recall that the equilibrium in the stage game requires one agent to perform the investment in co-operative skills, whereas the other agent can simply free-ride on that, by enjoying the bonus in joint production without paying the relative costs. In an evolutionary context, this argument entails that all the burdens and risks associated with the development of costly skills are carried out by only one of the two groups, and that social norms somehow sanctions this situation.

7.4 Social Norms and Uncertainty: The Possibility of a Mismatch Between Social Norms and Technology

7.4.1 Individual Risk-Aversion

In the above analysis, although the conception of 'objective' risk of competitive activities had been clearly identified from the outset and simply identified with the variance of an activity, the main limitation of the analysis consisted of the fact that agents were assumed to be risk-neutral. As a result, their choice of opting for the 'safer' co-operative activity, which we observed in some scenarios, was not due to their aversion towards the more risky activity, but it was a consequence of the lower expected payoff that more risky activities, by attaching heavy penalties to the occurrences of the 'bad' states of the world, i.e. being a loser in the market race, brought about. It still remains to be seen to what extent the introduction of risk aversion at the subjective level would change the picture. In this section, then, the

results of the investigation conducted in the case of subjective risk-aversion are reported.

The main result is that new scenarios arise, which show different patterns of inefficiencies from those already studied. Here, the individual utility function is assumed to have the following concave shape:

$$U(m) = m^{\phi} \tag{7.30}$$

where m is the argument of the utility function and ϕ is a coefficient smaller than one that measures the curvature of the utility function. Hence, the lower ϕ, the higher the individual risk aversion. The matrix of payoffs in Figure 7.5 obtains for the case of complementary efforts in co-operative production, when ϕ assumes a relatively low level, and the competitive technology is more efficient than the other. In particular: $\gamma = 0.8$; $\beta = 0.5$ [14]. It is apparent that $\alpha_i = 0$ is the dominant strategy for each agent, thus the outcome with no investment in co-operative skills will be selected.

	$\alpha_j=1$	$\alpha_j=0$
$\alpha_i=1$	-2.7 , -2.7	- 3.7 , 1.5
$\alpha_i=0$	1.5 , -3.7	1 , 1

Figure 7.5

What is perhaps surprising is that if the corresponding best reply function is analysed, then one finds that at the third stage the agents, rather than Racing, as it could be expected, decide to Co-operate. This is reported in (7.31)[15].

---

[14] More precisely, this is the set of parameter values that determine the payoffs in Figure 7.5: {p=10 ; $\gamma = .8$; $\beta = .5$; $\tau = 0.8$; $\rho = 0.8$; $\delta = 2$; $\phi = .1$}

[15] The values of the thresholds will not be reported, but, under the parameter values reported at the previous note, they assume a value strictly lying between 0 and 1.

$$s_i^*(\theta_i(\alpha_i),\nu_i(\alpha_i)) = \begin{cases} \begin{cases} C & \text{if } \alpha_j \le \bar{\alpha}_3 \\ R & \text{otherwise} \end{cases} & \text{if } \theta_i(\alpha_i)=1,\nu_i(\alpha_i)=1 \\ C & \text{if } \theta_i(\alpha_i)=1,\nu_i(\alpha_i)=0 \\ \begin{cases} C & \text{if } \alpha_j \le \bar{\alpha}_4 \\ R & \text{otherwise} \end{cases} & \text{if } \theta_i(\alpha_i)=0,\nu_i(\alpha_i)=1 \\ C & \text{if } \theta_i(\alpha_i)=0,\nu_i(\alpha_i)=0 \end{cases} \tag{7.31}$$

The first two cases and the last have a similar shape to those observed in (7.25), albeit this case represented the situation of perfect substitutes in individual effort. In particular, when $(\theta_i=1,\nu_i=1)$, the individual decide to Co-operate when the investment in co-operative skills is *not* too high. It is clear that this behaviour is led by the fear to compete with the other individual when he has comparatively good chances to win the contest. However, what differs with respect to both (7.25) and (7.22) is that this same behaviour is pursued by the individual when $(\theta_i=0,\nu_i=1)$: even when she is aware that the co-operative bonus is out of reach, the agent will choose to Co-operate if the other agent has a *low* enough level of investment in Co-operative skills.

It is clear that what determines such a behaviour is the individual's risk aversion, which prompts her to try to avoid any competition with the counterpart in all the situations in which he may look too likely to win the contest. Such a situation, then, determines a peculiar form of inefficiency under two different respects: first, the technology that is *ex-post- more* efficient, i.e. the competitive one, will never be selected by the individuals. Second, the individuals invest at the first stage in the skill that will never be used in the third stage. Notice, in fact, that if $\alpha_i=0$, then at the third stage one can be sure that Co-operation will be selected. The reason is that if they did not do so, they would be too vulnerable from the third stage in a possible race with the counterpart. Therefore, they both decide to invest in competitive skills. But this implies that, once arrived at the third stage, they both prefer to Co-operate in order to avoid to Race.

To be sure, the situation would change if the level of risk aversion decreased with respect to the previous value of ϕ. Still, this scenario appears a generic case for low enough values of this parameter. It highlights the existence of a peculiar type of

inefficiency between choices made at the two stages: although individuals are rational and apply backward induction in solving the game, they will still incur in this type of trap induced by risk aversion. This makes the present situation germane to that of radical uncertainty that is analysed in the next section.

7.4.2 Some Qualitative Considerations on Radical Risk

Up to now we have made use of a model in which 'uncertainty' was dealt with by means of probability measures that were known in advance and common knowledge between agents. Furthermore, even in the evolutionary parable, although the agent is initially pre-programmed to play a certain strategy, the type of actions she performs afterwards are consistent with what a rational player who is able to solve the game through backward induction would do. This is what the standard tools of Game Theory and theory of choice allow us to do in a formal way. However, the processes that affect technology matters are typically characterised by what can be called *radical* uncertainty (see Heiner (1983)). That is, not only some of the aspects that in the present setting were taken as certain, such as the technological parameters and the 'type', or ability, of the players, may be random variables, but also all of these may not be amenable to a representation through 'objective' probability measures. To be sure, one could defend standard rational choice in conditions of uncertainty by arguing for the construction of *subjective* measures of probability, but this would still be unsatisfactory as clearly common knowledge between subjective estimates would be required in order for a standard concept of solution to obtain. In other words, the agent may not be able to 'see through' to the final stages of the game, thus implying that she may not be able to solve the game by backward induction. Moreover, given the static nature of the basic setting, we have not been able to deal with the question of technical change. However, if we thought that technologies could change over time as an effect of either exogenous or endogenous processes, and that the choice of some technology was somehow irreversible, thus at least bringing about some costs in the option of changing it at some future stages, then the amount of uncertainty affecting the whole picture would be considerable. In the present section I wish to put forward some reflections on what such a situation of radical uncertainty and technical change could imply for the analysis and the results attained so far.

In the absence of social norms, arguably, by relying on an evolutionary argument we could still expect the system to converge to one of the equilibria existing in the stage game. The main reason would be that the process of trial and error, mutation and imitation, that characterises this process, could open the way to the discovering and exploitation of the benefits 'hidden' in the interaction, thus wiping away the behaviours that were not individually rational, at least in a long run perspective. That is, even if the technological parameters were unknown, and even if agents did not perform actions coherent with backward induction, the process of selection could still be relied upon in order to eradicate less successful strategies and make efficient ones thrive. Therefore, if the system was stuck in an outcome that was not a Nash equilibrium in the stage game, then it could be 'invaded' by some different behaviour that some 'mutant' agent would, sooner or later, perform. For instance, an agent that experimented with a 'competitive' way of behaviour in a context where all agents co-operated in a inefficient production system, would be able to reap extra payoffs, and her behaviour would, sooner or later, spread to the whole population. Admittedly, the convergence to the steady state may happen in the very long run, and it would not prevent the system to get stuck in *lock-in* traps. In particular, this would be the case if technological progress were subject to some forms of increasing returns to scale at the sectorial level and self-enforcing processes that we observed in Chapter 4 and 5. However, this would not undermine the scope of the previous analysis, as obviously inefficient steady states would be associated with Pareto-dominated equilibria in the stage game.

What I would like to argue is that, on the one hand, social norms may act as 'conservative' forces that hinder the process of experimentation of more successful strategies. In fact, we have seen how norms of behaviour, through acting on the motivational side of individuals, generally engender the self-reinforcement of outcomes that emerge as equilibria. This implies that when performing deviant behaviours, agents could take into account the resentment elicited in others by breaching some sort of established regularity of behaviour, especially when such a mutant behaviour inflicted some material costs on the opponents. This could indeed be the case in the example of the deviant competitor when co-operation is the

established behaviour. As a consequence, the process of elimination of unsuccessful strategies and convergence toward efficient social outcomes could be slowed down or prevented altogether, when the deviant behaviour is sanctioned with heavy costs associated with the breaching of normative expectations.

In fact, this argument would apply in particular in a context of radical uncertainty where co-operation was the general rule. In these conditions, carrying out a competitive strategy would immediately be seen as a 'hostile' strategy, independently of the actual payoffs that would be obtained at the end of the game. In fact, the action of investing in co-operative skill is non-negatively related with the opponent's payoff under any circumstance. In other words, although the actual structure of the interaction from the third stage could be somehow blurred to agents, they would be secure that the action of not investing in co-operative skills at the beginning by their opponent would certainly affect non-positively their own final result. In this situation, therefore, a deviant behaviour would be more likely to be deemed as detrimental to the interests of a co-operator, thus eliciting normative sanctions. Therefore, in these situations, agents could commit themselves to co-operative strategies even when an alternative technology was more 'efficient' in terms of aggregate production, and social norms could be the moral or normative sanction of this behaviour. All of these considerations would apply to the instance of rapidly developing technical change in the alternative technique with respect to that currently employed.

On the other hand, the analysis developed when dealing with the Prisoner's Dilemma scenario (sec. 6.3.2) shows the 'progressive' character that social norms may have. In that case, social norms were the main expedient through which an efficient outcome was engendered. Hence, social norms may at the same time be regarded as a form of 'social memory' for the society as a whole, in that they are the medium through which the success of past interactions can influence present situations and force agents to carry out efficient actions, although they clash with current self-interest.

CONCLUSIONS

The purpose of this thesis has been to investigate the bases for a comprehensive analysis of social norms and economic growth. Some foundational and substantive issues regarding the two topics have been tackled separately in the first and second part of the thesis respectively. Finally, an analysis embracing both aspects has been developed in the third part.

More precisely, aspects of the debate between moral philosophy and individual rational choice has been examined in Chapter 1, with the purpose of highlighting the underpinnings for a model of choice based on multiple motivations. By drawing on this analysis, a model has been developed in Chapter 2, where self-interested and other-regarding motivations combine in a *comprehensive* utility function. A particular specification in which 'social preferences' and 'intention-based' motivations are both considered is also developed. This elaborates on Rabin's model of fairness, its main idea being that individuals are *conditionally* willing to abide by their other-regarding motivation. In particular, the proviso upon which they condition their choice is the expectation of compliance by other agents with some shared ideas of the public interest. These ideas have found an application in Chapter 3, where the model of normative expectations, as implemented by Sugden, has been criticised on the grounds that other-regarding motivations are based on an 'empirical' notion of expectation, rather than on a 'causal' one. It has been argued that the lack of a substantive *reason to action* in conforming to the equilibrium may cause its 'instability' with respect to the introduction of some 'deviant' behaviour within the population of agents. In more general terms, a theory based on normative expectations only suffices to explain how a norm can be self-sustaining *supposing* it has already been established, but not how it has been brought about. In Chapter 4 a formal analysis of this argument has been developed by means of the dynamical investigation of the game, and it has been shown that this intuition receives support. The results of the model of conditional

compliance with public interest developed in the previous chapter have also been presented, suggesting that it may be deemed as a generalisation of Sugden's model.

In the second part of the thesis a model of growth has been developed, which is characterised by three key assumptions: bounded rationality of individuals, non-instantaneous market-clearing, and localized technical change at the technique level. The *multiple* steady states of the model have been analysed both analytically and numerically, and the conditions under which a lock-in to the slow-growth steady state have been spelled out. In particular, Chapter 5 and 6 have dealt with the cases of 'immobility' of labour amongst skill levels, and mobility up to some 'mobility' costs, respectively. These results have been interpreted as illustrating a different kind of 'poverty trap' than those put forward in the literature, in that markets forces do not suffice to *co-ordinate* agents on the efficient outcome. Some implications of political economy have also been advanced.

In Chapter 7 I have developed a model of 'institutional governance of uncertainty', where some of the relationships between technological choice, social norms and uncertainty have been investigated. This model relies on the above analysis in that agents act in accordance with the model of multiple motivations put forward in Part 1, and in that the evolution of the system and the occurrence of poverty traps is represented as in the growth model of part 2. The main idea has been that social norms can be seen as 'optimal' institutional designs to manage the risks involved with economic activities. In particular, social norms can either favour the undertaking of 'co-operative' as opposed to 'competitive' activities depending on the relative efficiency of the associated technologies. That is, when co-operative activities are more efficient than competitive ones from the aggregate point of view, but the associated risks would make competition the preferred strategy from the self-interested point of view, then social norms encouraging investment in co-operative skills can be relied upon to emerge, if agents attach sufficient importance to their other-regarding motivations. Analogously, when the competitive activity is more efficient, but risk-aversion by the individuals makes co-operation at the third stage the best-preferred option in the self-interested sense, then social norms eliciting competition can be expected to emerge. In any of these cases, social norms have a function analogous to

'social capital', or 'social memory' for a society, because they help to elicit the aggregate efficient outcomes, though this is contrary to self-interest. However, it has also been stressed that social norms could play a 'conservative' role by hampering the experimentation and exploration of possibly more efficient technological innovation that become available after the society has settled on a particular equilibrium. In this case, social norms themselves, by acting as a stabilising force in the perpetuation of an outcome, are one of the causes that determine the lock-in to a sup-optimal equilibrium.

Despite the fact that this model is highly stylised, I hope it may help to shed some light on some of the transformations that a number of countries are going through at present. In fact, through its emphasis on the necessity of a 'suitable match' between the institutional and economic framework within a society, the implications of the model for economic policy analysis is that some 'mismatch' between the social environment and economic policy can be deemed as being one of the causes of crisis in some countries. Japan is, I believe, an emblematic case with respect to this point. For many years the Japanese economy has been seen as stagnating in a 'trap' of low growth, low interest rates and surplus in current accounts. Although it is clear to many commentators that a vigorous policy of *institutional* reform, acting in particular on the corruption-prone and clan-inclined banking system, would be needed in order to break this vicious cycle, the question to be asked is why the same system fared so well up to only a decade ago. Is it that the competitive advantages of being a technological follower of the US eventually, and quite suddenly, were exhausted, or are there other more structural causes?

A different answer can perhaps be given if one looks at the process of concentration of economic activities in the service sector that Japan is going through[1]. The hypothesis could thereby be that the institutions that were – and still are - in place were well suited to the *past* structure of the economy, i.e. one that is highly concentrated in manufacturing, but are not well matched with the *current* structure of the economy. The reason for this interpretation is that manufacturing and services activities require different forms of 'optimal' risk management from the point of view

[1] See for instance Maddison (1982)

of the society as a whole. In particular, as manufacturing typically requires much larger fixed capital than services activity, the 'risks' associated with these activities, i.e. the economic losses in case of their unsuccessful termination, are higher. Consequently, they require a high 'institutional' coverage of risks, which, especially in Japan's situation of a country at the early stages of its development path[2], may have been better ensured by this type of banking system[3]. In other words, the huge investments that were needed to trigger off Japan's development could have perhaps been not undertaken if such inefficient banking system, which lowered the costs of risk through the *unconditional* insurance to bail out debts, was not in place. Therefore, having an *inefficient* credit system at the *market* level could instead be functional as an *efficient* management of risk at the *aggregate* level. In the light of the model developed in this thesis, this situation could be seen as a case in which the 'technological structure' of a society has moved *faster* than its social norms, thus causing a mismatch between them and contributing to bringing about economic stagnation.

Likewise, the debate on the reform of the welfare state and labour markets in continental Europe may be interpreted along the same lines. In this case, we can observe an evolution of the economic structure from the heavily manufacturing-based one that characterised the past decades to another one more concentrated on services, though still to a lesser degree than in the Anglo-Saxon countries. Therefore, these institutional reforms may be assessed on the grounds of the differences in risk-managing that they imply; on the one hand, the removal of welfare benefits and the process of making labour market more flexible could be seen as functional to the transition towards a 'weightless' economy such as the one that is concentrated on services, for the same reasons outlined in the case of Japan. On the other hand, this may prove not to be optimal in the long run, as significant risks may indeed be looming behind these 'new' activities, as the dramatic ups and downs of stock

[2] It has been argued (Acemoglu and Zilibotti (2001)) that risk is generally wider for countries at the *early* stages of development, as differentiation of economic activities can occur only with a lesser degree than more developed ones. Therefore, a co-operative set of institutions, rather than a competitive one, is likely to be optimal. Hence, this idea enhances the argument advanced above. *En passant*, notice that the policies advocated by the IMF and the World Bank for developing countries are generally encouraging, in a broad sense, more competition rather than more co-operation.

[3] For an institutional analysis germane to mine, see Aoki, from which many empirical aspects of the current argument have been drawn.

exchanges prices and the recent series of bankruptcies in the new-economy companies may indicate. In particular, although no large amount of physical capital is needed in order to start these activities, a significant amount of *human* capital may indeed be necessary, thus making this technology substantially similar to one based on manufacturing. As a consequence, the process of market deregulation that is occurring in these countries may be assessed with prudence.

Another case in which the present model could apply would be the transition to free-market and capitalistic management by formerly communist countries. In this case, the partially unsuccessful results that these reforms have obtained, may be accounted for in terms of the lack of appropriate type of social norms and attitudes of individuals in the face of significant structural changes in the economic sphere.

I hope these examples – sketchy though they are - suggest the relevance of an institutional analysis along the lines of the models developed in this thesis. To be sure, though, there is still considerable space for improvement and refinement of these concepts. First, the replicator dynamics has been assumed in the first part as the main rule for the evolution of individual behaviour. I have argued that this has to be understood as in the parable of the 'imitation of the most successful agent'. However, still are there some problems in identifying what 'success' means within a context that mixes material and ideological motivations as in the model of individual choice that has been adopted throughout. A criterion that paid attention to this distinction would represent significant progress in the theory.

Second, to some extent the reliance on other-regarding motivations in the account of individual behaviour and then social outcomes runs the risk of being seen as an *ad hoc* assumption. Although the formal analysis with which I have developed this model does impose *restrictions* on individual behaviour, it is true that the number of degrees of freedom in that model, including the normative function that makes up the other-regarding motivation, may simply be too many for a sound economic theory. Hence, more theoretical and empirical work in this area appears to be needed. An idea with respect to this point would be to turn our attention to the debate on the Humean model of choice and the relationship between beliefs and desires. Although in principle the traditional economic approach holds that desires have to be taken as

given, a 'rational' selection of desires may occur on 'empirical' grounds, by arguing that desires are more or less likely to be *untenable*, in failing to fulfil the consistency requirements of the rational choice theory. For instance, coming back to the well-known Humean example, preferring not to scratch one's finger over the end of the world, though possible in principle, would be hard to embed in a structure that did not incur violations of the transitivity principle when each of these two alternatives was matched with others.

As far as the growth model is concerned, more in-depth analysis would be needed on the type of technical progress that has been assumed, i.e. localised technical change, which adds to the assumption of increasing returns to scale at the technique level. Since the stark result of convergence to one of the two techniques depends on this assumption, further investigation of this aspect may improve the realism of the theory. In particular, if some boundaries on the productivity growth rates were assumed, as suggested by the idea of a 'life-cycle' of a technology, then *specialisation* rather than *convergence* could emerge from the model.

Finally, the model developed in the last part of the thesis is meant to depict some typical interactions within the socio-economic sphere that affect the aggregate system of risk-management. There would be an essentially similar relationship between workers in relation to a system of unemployment insurance, or between entrepreneurs and workers in relation to their activities in joint production and their general, as opposed to specific, investment. Although I believe that the general framework of this model can be relied upon to apply to these other relationships, undoubtedly a further deepening of the analysis is required. Moreover, I believe that the way 'risk' is modelled is still partially unsatisfactory, as this should typically be associated with the 'variance' of economic activities, rather than with lower expected payoffs deriving from uncertainty. However, the simple binomial distribution that has been taken on does not allow for this interpretation. Hence, an extension of the model with continuous random variables appears to be needed. Thirdly, ideally one may want to deal with the situation of radical uncertainty, but of course the lack of a viable model of bounded rationality makes this a really hard hurdle to surpass. Finally, attempting to apply the model to the practical cases mentioned above would be an

exciting exercise. In this sense, finding some correlations between the amount of risk involved in economic activities, their returns to scale, and the main structure of the economy, is a necessary step to progress along these lines.

Even with all these acknowledged limitations and required further developments, I still hope that the models put forward in this thesis, and the ideas that underpin these models, represent a viable framework to investigate the relationship between individual behaviour, social norms, and economic performance.

REFERENCES

- Acemoglu, D. and Zilibotti, F. (2001): "Productivity Differences", *Quarterly Journal of Economics*, vol. 116, pp. 563-606
- Aghion, P. and Bolton, P. (1997), "A Theory of Trickle-Down Growth and Development", *Review of Economic Studies*, vol. 64, pp. 151-172
- Aghion, P., Caroli, E., and GarcíaPeñalosa, C. (1999), "Inequality and economic growth: the perspective of the new growth theories", *Journal of Economic Literature*, vol. 37, pp. 1615-60
- Alesina, A. and Angeletos, G. (2002). *Fairness and Redistribution: US versus Europe* (mimeo)
- Anderlini, Luca and Antonella Ianni: *Path Dependance and Learning from Neighbours*, Games and Economic Behaviour 13: 141-77
- Anderson, P., Arrow, K. (Ed.), Pines, D., (Eds.), (1988). *The economy as an evolving complex system: Proceedings of the Evolutionary Paths of the Global Economy Workshop, held in September 1987, in Santa Fe*, Reading, Mass.; Wokingham: Addison-Wesley
- Andreoni, J. and Miller, J. (2000). "Giving according to the GARP: An Experimental Test of the Rationality of Altruism", *Mimeo*, University of Wisconsin and Carnegie Mellon University
- Andreoni, J. and Miller, J. (1993). "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence", *Economic Journal*, Vol. 103, pp. 570-585
- Andreoni, J. (1989). "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence", *Journal of Political Economy*, Vol. 97, pp. 1447-1458
- Andreoni, J. (1988). "Why Free Ride? Strategies and Learning in Public Goods Experiment", *Journal of Public Economics*, 37, 291-304
- Antonelli, C. (1995), *The Economics of Localized Technological Change and Industrial Dynamics*, Kluwer Academic Publisher

- Aoki, M. (2000). *Information, corporate governance, and institutional diversity : competitiveness in Japan, the USA, and the transitional economies*, Oxford : Oxford University Press
- Arrow, K. J. (1962), "The Economic Implications of Learning by Doing", *Review of Economic Studies*, vol. 29, pp. 155-173
- Arthur, W.B. (1985): *Competing Techniques and Lock-in by historical Events: The Dynamics of Allocation Under Increasing Returns*, Stanford: Stanford university Press
- Atkinson, A. B. and Stiglitz, J. E. (1969): "A New View of Technological Change", *Economic Journal*, vol. 79, pp: 573-8
- Axelrod, R. (1984). *The Evolution of Cooperation*, New York: Basic,
- Ayer, A. J. (1936), *Language, Truth and Logic.* Gollancz.
- Azariadis, C. and Drazen A. (1990). "Threshold externalities in Economic Development", *Quarterly Journal of Economics*, vol. 105, pp. 501-26
- Bacharach, M. (1999). "Interactive Team Reasoning: a Contribution to the Theory of Co-operation", *Research in Economics*, Vol. 53, pp. 117-47
- Banerjee, A. V. and Newman, A. F. (1998), "Information, the Dual Economy, and Development", *Review of Economic Studies*, vol. 65, pp. 631-653
- Barro R. J. (1991), "Economic growth in a Cross-Section of Countries", *Quarterly Journal of Economics*, vol. 106, pp. 407-44
- Barro, R. and Gordon, D. (1983). "A Positive Theory of Monetary Policy in a Natural Rate Model", *Journal of Political Economy*, Vol. 91, pp. 589 - 610
- Barro, R. and X. Sala-i-Martin (1995) *Economic Growth*, Boston, MA: Mc Graw Hill
- Barry, B. (1995). *A Treatise on Social Justice (Volume II): Justice as Impartiality*, Oxford: Clarendon Press
- Basu, S. e Weil, D. N. (1998), "Appropriate Technology and Growth", *Quarterly Journal of Economics*, vol. 113, pp. 1025-54
- Baumol, W. J. (1967). "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis", *American Economic Review*, vol. 57, pp. 415-26
- Ben Ner, Avner and Louis Putternam: "Values and institutions in economic analysis", in Avner Ben-Ner and Louis Putterman (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 3-69, 1998

- Benabou, R. (1996), "Equity and Efficiency in Human Capital Investment: The Local Connection", *Review of Economic Studies,* vol. 63, pp. 237-64
- Benaim, M. and Weibull, J. (2000). *Deterministic Approximations of Stochastic Evolution in Games*, mimeo
- Benhabib, J. and Rustichini, A. (1996), "Social Conflict and growth", *Journal of Economic Growth*, vol. 1, pp. 125-42
- Berg, J., Dickhaut, J. and McCabe, K. (1995). "Trust, Reciprocity and Social History", *Games and Economic Behavior*, Vol. 10, pp. 122-142
- Berman, E., J. Bound and Machin, S. (1998): "Implications of Skilled Biased Technological Change: International Evidence", *Quarterly Journal of Economics*, vol. 113, pp: 1245-1280
- Bernard, A. B. and Jones, C. I. (1996), "Technology and Convergence", *Economic Journal,* vol. 106, pp. 1037-44
- Bernheim, B. (1994): "A Theory of Conformity", *Journal of Political Economy*, Vol. 102, N. 5, pp.841-877
- Bicchieri, Cristina: *Norms of Cooperation*, in: Ethics 100 (July 1990), pp. 838-861
- Binmore, K. (2002), *Lecture and Handouts delivered at the Trento Summer School in Law and Economics*
- Binmore, K. (1999). "Why Experiment in Economics?", *Economic* Journal, V. 109, pp. F16-F24
- Binmore, K. (1998a). *Game Theory and the Social Contract: Just Playing*, Cambridge, MA: MIT Press
- Binmore, Kenneth (1998b): *A Utilitarian Theory of Political Legitimacy*, in Avner Ben-Ner and Louis Putterman (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 101-132
- Binmore, K., Gale, J. and Samuelson, L. (1995). "Learning to be Imperfect: The Ultimatum Game", *Games and Economic Behavior*, Vol. 8, pp.56-90
- Binmore, Kenneth (1993), *Game Theory and the Social Contract, Volume 1: Playing Fair*, Cambridge, Mass: MIT Press

- Blinder, A. and Choi, D.H. (1990): "A shred of evidence on theories of wage stickiness", *Quarterly Journal of Economics*, vol. 105, pp. 1003-1015
- Blount, S. (1995). "When Social Outcomes aren't Fair: The Effects of Causal Attributions on Preferences", *Organisational Behavior and Human Decision Processes*, 63, 131-144
- Bolton, G., Brandts, J. and Ockenfels, A. (1998). "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game", *Experimental Economics*, 3, 207-221
- Bolton, G. and Ockenfels, A. (2000). "A Theory of Equity, Reciprocity and Competition", *American Economic Review*, 100, 166-193
- Boyer, R. (ed.) (1997). *Contemporary Capitalism: The Embeddedness of Institutions*, Cambridge: Cambridge University Press
- Brandt, R. (1982). "Two Concepts of Utility", in Miller, H. and Williams, W. (eds.): *The Limits of Utilitarianism*, Minneapolis: University of Minneapolis Press
- Brennan, G. and Pettit, P. (2000). "The Hidden Economy of Esteem", *Economics and Philosophy*, Vol. 16, pp. 77 - 98
- Brewer, M.B. and Kramer, R.M. (1986). "Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing, *Journal of Personality and Social Psychology*, Vol. 50, pp. 543-549
- Brewer, M.B. and Gardener, W. (1996). "Who is the "we"? Levels of Collective Identity and self representation", *Journal of Personality and Social Psychology*, Vol. 71, pp. 83-93
- Brock, H. (1979): *"A Game theoretical Account of social Justice"*, in: Theory and Decision, 11, pp.239-265
- Broome, J. (1999). *Ethics out of Economics*, Cambridge: Cambridge University Press
- Buchanan, James: *The Limits of Liberty*, Chicago: University of Chicago Press, 1975
- Cameron, L. (1999). "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia", *Economic Inquiry*, Vol. 37 N.1, pp. 47-59
- Carpenter, J. (2000). "Punishing Free-Riders: The Role of Monitoring-Group Size, Second-Order Free-Riding and Coordination", *mimeo*, Middlebury College

- Charness, G. and Rabin, M. (2000). "Social Preferences: Some Simple Tests and a New Model", *mimeo*, University of California at Berkeley
- Cole, H., Mailath, G. and Postlewaite, A. (1998). "Class Systems and the Enforcement of Social Norms", *Journal of Public Economics*, Vol. 70, pp. 5 - 35
- Coleman, S.J.: *Foundations of Social Theory*, Cambridge, Mass: Harvard University Press, 1990
- Copp, David: *The Ring of Gyges: Overridingness and the Unity of Reason*, in: Social Philosophy and Policy, Cambridge University Press, Vol. 14 N.1, 1997
- Corradi, V. and A. Ianni (2000): "A simple locally interactive model of ergodic and nonergodic growth", *Southampton University Discussion Papers N. 0010*
- Cox, J. C. (2000). "Trust and Reciprocity: Implications of Game Triads and Social Contexts", *mimeo*, University of Arizona at Tucson
- Croson, R. (1999). "Theories of Altruism and Reciprocity: Evidence from Linear Public Good Games", Discussion Paper, Wharton School, University of Pennsylvania
- Cubitt, R. and Sugden, R., (2001). "On Money Pumps", *Games and Economic Behavior*, Vol. 37, pp. 121-60
- Dawes, R. M. and Thaler, R. (1988). "Cooperation", *Journal of Economic Perspectives*, Vol. 2, pp. 187-197
- Dosi, G. (1998), "Sources, Procedures and Microeconomic Effects of Innovations", *Journal of Economic Literature*, vol. 26, pp. 1120-71
- Dosi, G. and Nelson, R. (1993), "Evolutionary Theories in Economics", Laxenburg: International Institute for Applied Systems Analysis: *Working paper 9364*
- Dosi, G., Freeman, C., Nelson, R., and Soete, L. (1988), (eds.). *Technical Change and Economic Theory*, London: Pinter
- Dosi, G., Orsenigo, L. and Silverberg G. (1988), "Innovation, Diversity and Diffusion: A self-Organisation Model", *Economic Journal*, vol. 98, pp. 1032-1054
- Dufwenberg, M. and Kirchsteiger, G. (1998). "A Theory of Sequential Reciprocity", Discussion Paper, CentER, Tillburg University

- Durlauf, S. N. (1996), "A Theory of Persistent Income Inequality", *Journal of Economic Growth*, vol. 1, pp. 75-94
- Durlauf, S. N., Kourtellos, A. and Minkin, A. (2001) "The Local Solow Growth Model", *European Economic Review*, vol. 45, pp. 928-940
- Durlauf, S.N. (1993), "Nonergodic Growth", *Review of Economic Studies*, vol. 60, pp. 349-366
- Eichengreen, B. and T. Iversen, (1999). "Institutions and Economic Performance: Evidence from the Labour Market", *Oxford Review of Economic Policy*, Vol. 15, N. 4, pp. 121 - 138
- Elster, J. (1990). "Norms of Revenge", *Ethics*, V. 100, pp. 862-885
- Elster, J. (1989). *The Cement of Society,* Cambridge: Cambridge University Press
- Elster, J. (1986). "The Market and the Forum: Three Varieties of Political Theory", in Elster, J. and Hylland, A. (eds.), *Foundations of Social Choice Theory: Studies in Rationality and Social Change*, Cambridge: Cambridge University Press, 103-132
- Erev, I., and Roth, A. (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term", *Games and Economic Behavior*, Vol. 8, pp. 164-212
- Falk, A. and Fischbacher, U. (1999): *"A Theory of Reciprocity"*, Institute for Empirical Research in Economics, University of Zurich, WP N. 6
- Fagin, R., Halpern, J., Moses, Y. and Vardi, M.Y. (1995). *Reasoning About Knowledge*, Cambridge, MA: Mit press
- Falk, A., Fehr, E., and Fischbacher, U. (2000a). "Informal Sanctions", Institute for Empirical Research in Economics, University of Zurich, WP N. 59
- Falk, A., Fehr, E., and Fischbacher, U. (2000b). "Testing Theories of Fairness-Intentions Matter: ", Institute for Empirical Research in Economics, University of Zurich, WP N. 63
- Falk, A. and Gachter, S. (1999). "Reputation or Reciprocity?", Institute for Empirical Research in Economics, University of Zurich, WP N. 19

Fehr, Fischbacher, and Gachter (2001); "Are people conditionally cooperative? Evidence from a public goods experiment", *Economics letters*, 71, 397-404

- Fehr, E. and Fischbacher, U. (2000). "Third Party Punishment", *mimeo*, University of Zurich
- Fehr, E. and Gachter, S. (2000). "Cooperation and Pubishment in Public Goods Experiment", *American Economic Review*, 90, 980-994
- Fehr, E., Kirchsteiger, G., and Riedel, A. (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation", *Quarterly Journal of Economics*, Vol. 108, pp. 437-460
- Fehr, E. and Schmidt, K. (2001): *"Theories of Fairness and Reciprocity – Evidence and Economic Applications"*, Institute for Empirical Research in Economics, University of Zurich, WP N. 75
- Fehr, E. and Schmidt, K. (1999): "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics*, Vol. 114, pp. 817-868
- Fehr, E. and Tougareva, E. (1995). "Do High Monetary Stakes Remove reciprocal Fairness? Experimental Evidence from Russia", *Mimeo*, Institute for Empirical Research in Economics, University of Zurich
- Fershtman, C. and Weiss, Y. (1998). "Why do we care what others think about us?", in Avner Ben-Ner and Louis Putterman (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 133-50
- Forsythe, R., Horowitz, J., Savin, N., and Sefton, M. (1994). "Fairness in Simple Bargaining Games", *Games and Economic Behavior*, Vol. 6, pp. 347-369
- Frank, Robert: *Passions within Reason*, New York: W.W. Norton & C., 1988
- Frankel, D. and Pauzner, E. (2000). "Resolving Indeterminacy in Dynamic Settings: The Role of Shocks", *American Economic Review*, Vol. 115 N. 1, pp. 285-304
- Friedman, D. (1998). *"Evolutionary Economics Goes Mainstream: a Review of the Theory of Learning in Games"*, Journal of evolutionary economics, V. 8, pp.423-32
- Fudenberg, D. and Levine, D. (1998). *The Theory of Learning in Games*, Cambridge (MA): MIT Press
- Galor, O. (1996). "Convergence? Inferences from Theoretical Models", *Economic Journal*, 106, 1056-1069

- Galor, O. and Weil, D. (1996), "The Gender Gap, Fertility, and Growth", *American Economic Review*, vol. 86
- Galor, O. and Zeira, J. (1993), "Income Distribution and Macroeconomics", *Review of Economic Studies*, vol. 60, pp. 35-52
- Gauthier, D., (1986). *Morals by Agreement*, Oxford: Oxford University Press
- Geanakoplos, John, Pearce David, and Stacchetti Ennio: *Psychological Games and Sequential Rationality*, in: Games and Economic Behavior 1, 60-79 (1989)
- Gilbert, M. (1989). *On Social Facts*, London: Routledge
- Goodwin R M (1967), "A Growth Cycle", in: Feinstein, CH (ed): *Socialism, Capitalism and Economic Growth*, London: Mac Millan
- Greif A. *et al.* (2002). "Institutions and impersonal exchange: from communal to individual responsibility", *Journal of institutional and theoretical economics*, Vol. 158, pp. 168-218
- Greif, A. (1994). "Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies", *Journal of Political Economy*, Vol. 102, pp. 912 - 950
- Griffin, J. (1977). "Are There Incommensurable Values?", *Philosphy and Public Affairs*, V. 7, pp. 39-59
- Griffin, J. (1986). *Well-Being*, Oxford: Clarendon Press
- Griffin, J. (1991): "Against the Taste Model", in Elster, J. and Roemer, J.E. (eds.) *Interpersonal Comparisons of Well-being*, Cambridge: Cambridge University Press, pp. 45-69 also in: Alan Hamlin (ed.): *Ethics and Economics*, Cheltenam: Elgar, 1996
- Grimalda, G. (2001), *Crescita con tecnologie in competizione: un approccio evolutivo* (translation: *Growth with Competing Technologies: An Evolutionary Approach),* Pavia: Ph.D dissertation
- Grimalda, G. and Sacconi, L. (2002). "The Constitution of the Nonprofit Enterprise: Ideals, Conformism and Reciprocity", *LIUC Papers n. 115*
- Grether, D.M. and Plott, C.R. (1979). "Economic Theory of Choice and the Preference Reversal Phenomenon", *American Economic Review*, Vol. 69, pp. 623-638

- Grossman, S. and Hart, O. (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration", *Journal of Political Economy*, Vol. 94, pp. 691-719
- Guckhenheimer, J. and Holmes, P. (1990): *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Berlin: Springer Verlag
- Guth, W. and van Damme, E. (1998). "Information, Strategic Behavior and Fairness in Ultimatum Bargaining: an Experimental Study", *Journal of Mathematical Psychology*, Vol. 42, pp. 227-247
- Guth, W., Kliemt, H., and Ockenfels, A. (1982). "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organisation*, Vol. 3, pp. 367-388
- Hardin, Russell: *Morality within the Limits of Reason*, Chicago: Chicago University Press, 1988
- Hargreaves-Heap, S., and Varoufakis, Y. (2002), "Some Experimental Evidence on the Evolution of Discrimination, Co-operation and Perceptions of Fairness", *Economic Journal*, V. 112, pp. 679-703
- Hargreaves-Heap, S., Hollis, M., Lyons, B., Sugden, R., Weale, A., (1992). *The Theory of Choice: A Critical Guide*, Oxford: Blackwell
- Hare, R. (1982). *Moral Thinking: its Levels, Method and Point*, Oxford: Oxford University Press
- Hare, R. (1952). *The Language of Morals*, Oxford: Oxford University Press
- Harsanyi, J. (1969). "Rational Choice Models of Behaviour versus Functionalist and Conformist Theories", *World Politics*, Vol. 22, pp. 513-38
- Harsanyi, J. (1977): *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge: Cambridge Unibersity Press
- Heiner, R. (1983). "The Origins of Predictable Behavior", *American Economic Review*, Vol. 73, pp. 560-595
- Hildebrand, F. B. (1987): *Introduction to Numerical Analysis*, Dover Books on Advanced Mathematics
- Hill, T. (1997). "Reasonable Self-Interest", *Social Philosophy and Policy*, Cambridge: Cambridge University Press, Vol. 14, pp. 52-85

- Hirsch M.W. and Smale S. (1974): *Differential Equations, Dynamical Systems, and linear Algebra*, London: Academic Press
- Hoffman, E., McCabe, K., and Smith, V. (1996). "On Expectations and Monetary Stakes in Ultimatum Game", *International Journal of Game Theory*, Vol. 25, pp. 289-301
- Hoffman, E., McCabe, K., Scachat, K. and Smith, V. (1994). "Preferences, Property Right, and Anonymity in Bargaining Games", *Games and Economic Behavior*, 7, 346-380
- Hodgson. G. M. (1999), *Evolution and institutions: on evolutionary economics and the evolution of economics*, Northampton, MA: Edward Elgar Publishing
- Hogarth, R., and M. Reder (eds.), (1986). *Rational Choice*, Chicago: The University of Chicago Press
- Hollis, M. (1998). *Trust Within Reason*, Cambridge: Cambridge University Press
- Holland, F. Watson, F., Wilkinson, J. (1974). *Introduction to Process Economics*
- Hume, David (1740). *A Treatise of Human Nature*, Oxford: Clarendon Press (reprinted 1978)
- Hume, David (1777). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, Oxford: Clarendon Press (reprinted 1998)
- Isaac, R., McCue, K., and Plott, C. (1985). "Public Good Provision in an Experimental Environment", *Journal of Public Economics*, 26, 51-74
- Kagel, J. and Roth, A. (eds.), (1995). *Handbook of Experimental Economics*, Princeton: Princeton University Press
- Kahneman, D., Knetsch, J., and Thaler, R. (1986). "Fairness and the Assumptions of Economics", *Journal of Business*, 59, S285-S300
- Kandori, M. (1992). "Social norms and Community Enforcement", *Review of Economic Studies*, Vol. 59, pp. 63 - 80
- Kapur, B. K. (1999). "A Communitarian Utility Function and Its Social and Economic Implications", *Economics and Philosophy*, Vol. 15, pp. 43 - 62
- Kolpin, V. (1992): "Equilibrium Refinements in Psychological Games", *Games and Economic Behavior*, Vol. 4 N. 2, p. 218-228

- Kreps, D., Milgrom, P., Roberts, J., and Wilson, R. (1982). "Rational Cooperation in Finitely Repeated Prisoners' Dilemma", *Journal of Economic Theory*, 27, 245-252
- Krugman, P. (1991). "History Versus Expectations", *Quarterly Journal of Economics*, 106, 651-667
- Le Caldano, E. (1991). *Hume e la nascita dell'etica contemporanea*, Bari: La Terza
- Levine, D. (1998). "Modelling Altruism and Spitefulness in Experiments", *Review of Economic Dynamics*, Vol. 1, pp. 593-622
- Lewis, David (1969): *Convention: A Philosophical study*, Cambridge, Massachusetts: Harvard University Press,
- Lindbeck, A., Nyberg, S., and Weibull, J. (1999). "Social Norms and Economic Incentives in the Welfare State", *Quarterly Journal of Economics,* V. 114, pp.1-35
- Lucas, R. E. (1988), "On the Mechanics of Economic Development", *Journal of Monetary Economics*, vol. 22, pp. 3-42
- Lucas, R. (1986)."Adaptive Behavior and Economic theory", in: Hogarth, R., and M. Reder (eds.). *Rational Choice*, Chicago: The University of Chicago Press
- Mailath, G. (1998). "Do People Play Nash Equilibrium? Lessons from Evolutionary Game Theory", *Journal of Economic Literature*, 36, 1347-1374
- Mainard Smith, J. and G.R. Price: *The Logic of Animal Conflict*, in. Nature, 246: 15-18, 1973
- Mainard Smith, J.: *Evolution and the Theory of Games*, Cambridge: Cambridge University Press, 1982
- Mankiw, N., Romer, D., and D. N. Weil (1992), "A Contribution to the Empirics of Economic Growth", *Quarterly Journal of Economics*, vol. 107, pp. 407-37
- Margolis, Howard: *Equilibrium Norms*, in: Ethics, 100 (July 1990), pp, 821-837
- Marschak, J. and Radner, R. (1972). *Economic Theory of Teams*, New Haven: Yale University Press
- Marwell, G. and Ames, R. (1981). "Economists Free Ride, Does Anyone Else?", *Journal of Public Economics*, 15, 295-310
- Matsuyama, K. (1991). "Increasing Returns, Industrialization, and Indeterminacy of Equilibrium", *Quarterly Journal of Economics*, 106, 617-650

- Maurseth, P. B. and Verspagen, B. (1999), "Europe: one or several systems of innovation? An analysis based on patent citations", in: Fagerberg, J., Guerrieri, P. and Verspagen, B. (eds.): *The Economic Challenge for Europe*, Cheltenam (UK): Edward Elgar
- Mertens, J. F. and Zamir, S. (1985), "Formulation of Bayesian analysis for games with incomplete information", *International Journal of Game Theory*, Vol. 14, No. 1, pp. 1-29
- McCabe, K., Rigdon, M. and Smith, V. (2000). "Positive Reciprocity and Intentions in Trust Games", *mimeo*, University of Arizona at Tucson
- Moav, O., (2002) "Income distribution and macroeconomics: the persistence of inequality in a convex technology framework", *Economic Letters*, vol. 75, pp. 187-92
- Murphy, K. M., Shleifer, A., and Vishny R. W., (1989), "Industrialization and the Big Push", *Journal of Political Economy*, vol. 97, pp. 1003-1026
- Myerson, R; (1991), *Game Theory: Analysis of Conflict*, Cambridge, MA: Harvard University Press
- Nelson R, Winter S. G. (1982), *An Evolutionary Theory of Economic Change*, Cambridge MA: The Belknap Press of Harvard Univesity
- Nelson, R. and S. Winter, (1982): *An Evolutionary Theory of Economic Change*, Cambridge MA: Harvard University Press
- North, D.C. (1990): *Institutions, Institutional Change and Economic Performance*, Cambridge MA: Cambridge University Press
- Nozick, R. (1975), *Anarchy, state and Utopia*, Oxford: Blackwell
- OECD: *Science, Industry and Technology: a Scoreboard of Indicators*, 1997
- Ockenfels, A. and Selten, R. (1998). "An Experimental Solidarity Game", *Journal of Economic Behavior and Organisation*, 34, 517-539
- Offerman, T., Schram, A. and Sonnemans, J. (1999). "Strategic Behavior in Public Good Games – When Partners Drift Apart", *Economics Letters*, 62, 35-41
- Okuno-Fujiwara, M. and Postlewaite, A. (1995). "Social Norms and Random Matching Games", *Games and Economic Behavior*, Vol. 9, pp. 79 - 109

- Okuno-Fujiwara, M., Prasnikar, V., Roth, A. E. and Zamir, S. (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study", *American economic Review*, Vol. 81, pp. 1068-95
- Oswald, A. and Zizzo, D. (2000). "Are People Willing to Pay to Reduce Others' Income", *mimeo*, Oxford University
- Parente, S. L. and Prescott, E. C., (1994), "Barriers to Technology Adoption and Development", *Journal of Political Economy*, vol. 102, pp. 298-321
- Persson, T. and Tabellini, G. (1994), "Is Inequality Harmful for Growth? Theory and Evidence", *American Economic Review*, vol. 84, pp. 600-21
- Persson, T. And Tabellini, G. (2000). *Political Economics: Explaining Economic Policy*, Cambridge (MA): MIT Press
- Pettit, P. (1990). "Virtus Normativa", *Ethics*, Vol. 100, pp. 725-755
- Pettit, P. (1993). (ed.) "Introduction", in: *Consequentialism - The International Research Library of Philosophy*, Vol. 6, Aldershot: Dartmouth
- Pettit, Philip and Robert Sugden: *The Backward Induction Paradox*, Journal of Philosophy 86: 169-83, 1989b
- Pettit, Philip: *Free Riding and Foul Dealing*, Australian Journal of Philosophy 67, 1989a
- Quah, D. (1996), "Twin Peaks: Growth and Convergence in Models of Distribution Dynamics", *Economic Journal*, 106, 1045-1055
- Quah, D. (1997), Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs, *Journal of Economic Growth*, vol. 2, pp. 27-59
- Rabin, M. (1993): "Incorporating Fairness into Game Theory", *American Economic Review*, Vol. 83, N. 5, pp. 1281-1302.
- Rabin, M. (1995), *Moral Preferences, Moral Constraints, and Self-Serving Bias*, Working Paper, Department of Economics, University of California, Berkeley

Rawls, J, (1996). *Political Liberalism*, New York; Chichester: Columbia University Press

- Rawls, J. (1971), *A Theory of Justice*, Oxford: Clarendon Press
- Roback Morse, J. (1997). "Who is Rational Economic Man?", *Social Philosophy and Policy*, Cambridge: Cambridge University Press, Vol. 14,

- Roback Morse, Jennifer: *Who is Rational Economic Man?* in: Social Philosophy and Policy, Cambridge University Press, Vol. 14 N.1, pp. 1997
- Romer, P. M. (1993), "Idea Gaps and Object Gaps in Economic Development", *Journal of Monetary Economics*, vol. 32, pp. 543-72
- Romer, P.M. (1994), "The Origins of Endogenous Growth", *Journal of Economic Perspectives*, vol. 8, pp. 3-22
- Rose-Ackerman, S. (1996).: "Altruism, Nonprofits, and Economic Theory", *Journal of Economic Literature*, Vol. 34, N. 2, pp. 701-728
- Rosenberg, N. (1982): *Inside the Black Box: Technology and Economics*, Cambridge: Cambridge University Press
- Rubinstein, A. (1998). *Modeling bounded rationality* Cambridge, Mass.; London : MIT Press
- Sacco, Pier Luigi: *On the dynamics of social norms*, in: Cristina Bicchieri, Richard Jeffrey and Brian Skyrms (eds.): *The dynamics of norms* Cambridge: Cambridge University Press, Ch. 3, 1997
- Sacconi L. (2002); *The Efficiency of the Non-profit Enterprise: Constitutional Ideology, Conformist Preferences and Reputation,* Liuc Papers n. 110
- Sacconi, L. (2000), *The Social Contract of the Firm*, Berlin: Springer
- Sacconi, Lorenzo: *Eduzione vs. evoluzione nella selezione dell'equilibrio: un'alternativa all'analisi dell'insorgenza dell'ordine in Hayek*, Quaderni di Storia dell'Economia, IV/1986/3, pp. 157-201
- Salter, W. E. G (1969), *Productivity and Technological Change*, Cambridge: Cambridge University Press
- Sargent, T. (1993). *Bounded rationality in macroeconomics : the Arne Ryde memorial lectures* Oxford: Clarendon Press
- Scanlon, T. M. (2001). "Symposium on Amartya Sen's Philosophy: 3 Sen and Consequentialism", *Economics and Philosophy*, Vol. 17, pp. 39 - 50
- Schelling, Thomas C.: *Strategy of Conflict,* Cambridge, Mass.: Harvard University Press, 1960

- Schmidz, David: *Rational Choice and Moral Agency*, Princeton: Princeton university Press, 1995
- Segal, U. and Sobel, J. (1999). "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings", *Mimeo*, University of California at San Diego
- Sen, A. (1979): *Interpersonal Comparisons of welfare"*, in M.J. Boskin (ed.), *Economics and Human Welfare*, New York: Academic press, 183-201
- Sen, A. (1985), 'Well-being, Agency and Freedom', *Journal of Philosophy*, 82: 169-221
- Sen, A. (2000), 'Consequential Evaluation and Practical Reason'. *Journal of Philosophy*, 97: 477-502
- Sen, A. (2001). "Symposium on Amartya Sen's Philosophy: 4 Reply", *Economics and Philosophy*, Vol. 17, pp. 51 - 66
- Silverberg G. and Verspagen, B. (1994), "Collective Learning, Innovation and Growth in a Boundedly Rational, Evolutionary World", *Journal of Evolutionary Economics*, vol, 4, pp. 207 - 26
- Simon, H. (1955). "Behavioral Model of Rational Choice", *Quarterly Journal of Economics*, V.69, pp 99-118
- Skyrms, Brian: *Chaos and the explanatory significance of equilibrium: Strange attractors in evolutionary game dynamics*, in: Cristina Bicchieri, Richard Jeffrey and Brian Skyrms (eds.): *The dynamics of norms* Cambridge: Cambridge University Press, Ch. 10, 1997
- Smith, A. (1759): *The Theory of Moral Sentiments*, Oxford: Clarendon Press, (reprinted: 1976)
- Smith, M. (1995). *The Moral Problem*, Oxford (UK) and Cambridge (MA): Blackwell
- Stiglitz, J. E. (2003). *Globalization and its discontents*, London: Penguin
- Sugden, R. (2000). "Team Preferences", *Economics and Philosophy*, Vol. 16: 175-204
- Sugden, R. (1998a): *The motivating power of expectations*, mimeo; published in: "The Motivating Power of Expectations", in J. Nida-Rumelin and W. Spohn, (eds). *Rationality, Rules and Structure*, Amsterdam: Kluwer, pp. 103-29
- Sugden, R. (1998b). *Normative expectations: the simultaneous evolution of institutions and norms*, in Avner Ben-Ner and Louis Putterman (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 73-100

- Sugden, R. (1993). "Thinking as a team: Towards an explanation of nonselfish behavior", *Social Philosophy and Policy*, Vol. 10, pp. 69-89
- Sugden, R. (1990). "Contractarianism and Norms", *Ethics*, Vol. 100, pp, 768-786
- Sugden, R. (1986). *The Economics of Welfare, Rights and Co-operation*, Oxford: Basil Backwell
- Thaler, (1988). "The Ultimatum Game", *Journal of Economic Perspectives*, Vol. 2, 195-206
- Taylor, M. (1987). *The Possibility of Cooperation*, Cambridge: Cambridge University Press
- The Economist (2002), *Doubts inside the barricades - The IMF*, 28 September 2002, p.90
- Tuomela, R. (1995). *The Importance of Us*, Stanford, CA: Stanford university Press.
- Turner, J. and Giles, H. (1981). *Intergroup Behavior*, Chicago: University of Chicago Press
- Ulmann-Margalit, Edna: *Revision of Norms* Ethics, 100 (July 1990), pp, 756-767
- Ulmann-Margalit, Edna: *The Emergence of Norms*, Oxford: Oxford University Press, 1977
- van de Kragt, A., Orbell, J., and Dawes, R. (1988). "Explaining Discussion Induced Cooperation", *Journal of Personality and Social Psychology*, 54, 811-819
- van de Kragt, A., Orbell, J., and Dawes, R. (1983). "The Minimal Contributing Set as a Solution to Public Goords Problems", *American Political Science Review*, 77, 112-22
- Veblen, T. (1922). *The Theory of the Leisure Class – An Economic Study of Institutions*, London: George Allen Unwin (first published 1899)
- Verbeek, B. (2001), *Consequentialism, Rationality and the Relevant Description of Outcomes*, Economics and Philosophy, V. 17, pp. 181 - 205
- Verspagen, B. (1993), *Uneven Growth between Interdependent Economies*, Aldershot: Avebury
- Vivarelli, M. (1995), *The Economics of Technology and Employment*, Cheltenham: Edward Elgar
- von Hayek, F.: *Law, Legislation and Liberty,* vol. 1 *Rules and Order*, London: Routledge &Kegan Paul, 1973

- Weibull, J. (1995), *Evolutionary Game Theory*, Cambridge: Mit Press
- Weibull, Jorgen W.: *Evolutionary Game Theory*, Cambridge, Mass: The MIT Press, 1995
- Wiggins, D. (1991). *Needs, Values, Truth*, Oxford (UK) and Cambridge (MA): Blackwell
- Williams, B. (1972). *Morality*, Cambridge: Cambridge University Press
- Williamson, J. (2002). *Did the Washington Consensus Fail?*, Outline of Remarks at the Center for Strategic & International Studies
- Young, Peyton H., (1998). *Individual Strategy and Social Structure*, Princeton: Princeton University Press
- Young, Peyton H (1993). "The Evolution of Conventions", *Econometrica*, 61: 57-84