Funk Games
-approaching collective rationality-

by
Jyrki Konkka

***Der Einwand, der Seitensprung, das Fröhliche, Mißtrauen, die Spottlust sind Anzeichen der Gesundheit: alles Unbedingte gehört in die Pathologie*. (§154)**
Friedrich Nietzsche (1896): *Jenseits von Gut und Böse*.

ISBN 952-91-2109-1 (Nid.)
ISBN 952-91-2110-5 (PDF version)
Yliopistopaino
Helsinki 2000

Acknowledgments

This essay is about collective rationality. Thus, the issue is a conjoint of two distinct sources of human action, namely rationality and sociality, which are often thought to conflict with each other. The standpoint of conflict between the two sources is rejected here. Instead, an idea that individual rationality is an essential part of the emergence of collective action and an idea that collective expectations in part generate rational action is embraced in this essay.

The starting point of this study is in the mid- 1980's, when Professor Martti Kuokkanen introduced me the problems of free riding, collective action and rationality. Professor Raimo Tuomela encouraged me to continue with these issues in early 1990's, and Professor Matti Sintonen supported me in several ways during concluding this effort. I also wish to acknowledge Professor Jack Vromen, who gave me the possibility to visit in Erasmus Universiteit Rotterdam to present some early versions of this study. Of course, all the years of teaching, studying and staying at the Department of Social and Moral Philosophy in University of Helsinki have credited to this study.

Writing this dissertation provided me a wonderful journey through the fields of philosophy. However, all the material I ran through during this journey is not included in this essay. The main reason for leaving some material outside was to provide a coherent picture of what I thought to be the most essential questions concerning collective rationality. Thus, for instance, the questions of moral philosophy and axiology are put aside for now. Instead, the focus is on methodological, social and mental philosophical, issues concerning human behavior. Needless to say, there is only a thin blue line separating the different philosophical questions from each other.

I wish to thank everyone who has aided me in any way to conclude this work. I am grateful to Professor Martti Kuokkanen and Professor Gabriel Sandu for reading, commenting and criticizing the penultimate version of the manuscript. Professor Raimo Tuomela, Dr. Olli Loukola and Mr. Pekka Mäkelä read and made some helpful comments on the manuscript, as well. Professor Henry Fullenwider revised my English for which reason I am most grateful to him.

I am also grateful to those Foundations whose financial aid made the beginning of this project possible. They are: Alfred Kordelin Foundation, Finnish Cultural Foundation, Emil Aaltonen Foundation and E. J. Sariola Foundation. Academy of Finland and University of Helsinki have financed my journeys to symposia and congresses abroad.

Finally, special thanks are to be expressed to my mother Anita Konkka my father Esko Pasanen and his spouse Marja, and to my spouse Riikka Perälä whom I dedicate this book.

Helsinki, April 20, 2000

Jyrki Konkka

Abstract

The study focuses on collective rationality in the behavior of rational agents. The starting-point is, however, individual rationality as a guiding and explanatory principle from the point of view of the optimal use of resources. For instance, the model of maximizing expected utility (= EU) provides the required basis. However, the notion of collective rationality does not reduce without a residual to the notion of maximizing the EU. If it did, the introduction of collective rationality would be epiphenomenal: redundant in other than a clarifying sense. Closer examination, however, reveals that collective rationality has certain irreducible functions. For instance, it corrects the problems of unsatisfactory and indeterminate resolutions that would confront rational individuals in separation. Hence, although the explanatory strategy of the study is in part reductive, its basic ideology is essentially non-reductive. This ambivalent feature of the study is reasonable in virtue of its subject matter.

Collective rationality is expressed in terms of social habits, practices and norms. I will show that the standard game theoretical notions are unable to explicate some forms of rational social behavior, and they do not as such capture the collective rationality expressed by the social habits. Especially, the notion of rational choice does not correctly describe the behavior of following norms, practices and rules. Instead, the corresponding behavior is better described in terms of conformity to, and mutually expected conformity to, a given practice. The standard solution concept, equilibrium, is replaced by the notion of 'path'. Instead of making choices that enforce equilibria agents conform to a prevailing practice, and this conformity enforces a path that in turn yields suitable expectations for the future course of action. The conformity and the mutual expectations for conformity describe a social interaction situation that I will call the 'FUNK GAME'. In funk games, and in collective rationality in general, the emphasis is on the indirect maximization of the EU. Rational agents aim at satisfying mutual expectations, and above all at coordinating their actions, even though the maximization of the EU is at the core of rational behavior. The emphasis on collective rationality introduces the important constraint that the skills and capacities of the agents are limited. Especially, rationality is essentially incomplete.

A new account is given of collective action and its problems. Free riding is now satisfactorily explicated from the point of view of rational agents. The notion of the funk game makes it possible to see individual and collective rationality in coexistence with each other.

Key words: rationality, collective action, social habit, free riding, funk game

CONTENTS

1 INTRODUCTION

According to a common phrase, man is a rational animal. Another maxim has it that man is a political animal. Although man can be described by an untold number of attributes I shall concentrate mainly on these two in a discussion of *Homo economicus sociologicus*. However, I should first specify what I understand by this description.

I take 'sociologicus' to mean, especially, that man is a social animal capable of collective action.[1] The term 'economicus' signifies that man is an intelligent animal capable of the optimal use of his resources. Together, these attributes point out the main topic of this essay: collective rationality. But that is not all. Since I am dealing with human beings as agents, I take the limitations of their skills and capacities into account. Their rationality is essentially incomplete. But what is lost in skill and capacity is compensated for by cooperative action: by means of collective rationality, certain shortcomings of individual rationality can be resolved. However, a fertile collision of individual heads cannot be accounted for just like that. Collective rationality is not just a mereological sum of the individually best choices. Interdependence of action and witnessed success in coordination and cooperation in part determine the collectively rational thing to do. Thus, the *emphasis* in this essay is on collective rationality. Even so, however, the *standpoint* is initially and ultimately that of an individual agent. After all, collective rationality is implemented in individual actions, expectations and evaluations, or in individual reasoning, although it cannot be reduced to the properties of an individual mind. Relatedness and factuality are important factors in determining the collectively expected option in a course of action.

As suggested above, the constituting factors of the subject matter of this essay can be organized into following schema. In this study, rationality as optimal use of resources refers to 'economy of mind', and rational behavior is often defined as maximization of expected utility. Despite its apparent shortcomings, the idea of maximizing expected utility provides a good approximation of truthfully accounting for rational behavior, not least because it is a well-defined approach. However, for certain clarifying purposes I shall apply the notion of maximizing expected utility to describe a certain type of rational behavior that is not, strictly speaking, maximizing expected utility, although the function of the corresponding behavior may be that. The main purpose for using the notion a bit misleadingly is to remind the reader that rationality is understood basically as individual optimal use of resources.

I make a distinction between direct and indirect maximization of expected utility. The former is meant to describe individual rationality and the latter collective rationality. Collective rationality is not implemented by directly aiming at instant maximization, but by conforming to certain mutually expected practices whose function is to maximize the expected utility. This function may be, and generally is, latent in the present course of action. Still, conforming to the corresponding practice is considered to be rational behavior. Emphasizing conformity points out that rational behavior does not always require present deliberation. Especially, even if the behavior does not aim at direct maximization, it may still be considered as an optimal use of the resources in the required sense. Thus, the standard notion of rationality, which is embraced e.g. by game theory and described in terms of direct maximization of expected utility, is inadequate. Parallel to making a deliberated choice from among alternative options, rational behavior can also be implemented in a non-reconsiderated manner by conforming to a salient option. This means, especially, that the standard conception of a game is insufficient to describe rational behavior under certain recurring social interaction situations.

Social conventions, customs, norms and institutions are not in general deliberately picked out from several options, and following e.g. a social norm is rarely a matter of a given, presently deliberated rational choice. Rather, their realization is based on conformity to an existing path and to the mutual expectations of virtually everyone conforming to this path. Still, the corresponding behavior is under normal conditions considered rational. Especially, 'path' can be substituted for 'equilibrium' as the solution concept, and the strategic move of an agent is 'conformity' rather than 'choice'. This construct constitutes a type of social interaction situation which I call the 'funk game'.

Funk games are games of collective rationality. They consist of moves which conform to prevailing practices in the sense that this conformity implements a path, and conformity to this path brings about a resolution of the corresponding social dilemma. The resolution that implements the collectively rational option I call, maybe a bit artificially, the 'social habit'. Thus, social conventions, customs, norms and institutions are all subsumed under the notion of social habit. The social habit expresses the existence of funk games.

Although conformity and the corresponding expectations constitute funk games, the element of deliberation is not by-passed. After all, it is an important element of individual rationality, and collective rationality does not prevent acts of individual rationality even if the two appear to be in conflict. Especially, it is individually rational to take into account general conformity and mutual expectations about this conformity when one finds oneself in the situation of reconsidering the reasonableness of a

[1] Here, when I use the term 'man' I mean, naturally, human being. No sexist connotations should be sought here or

prevailing course of action and its alternatives. Sometimes it is rational to choose against the general conformity. The main point here is that present deliberation, which aims at the direct maximization of expected utilities, may nest in funk games. For instance, this is the case when free riding occurs.

***

The study of free riding and the problem of collective action have provided the initial motivation for this work, and funk games provide the answer to the initial research problem stated by that study. A detailed discussion of problems of collective action and free riding is presented in Chapter 2, and the discussion that anticipates the funk games is presented in the following chapters. The basic idea of introducing funk games came as a result of the observation that standard game theoretical analysis appears to be inadequate for describing e.g. the free rider problem. The standard models do not leave any room for expectations concerning a free ride, either because it is in everyone's interest to defect or because coordination of action predominates over the current course of action. Furthermore, the standard models misconceive the essential features of the problems of collective action. The aspect of coordination is factually left out of the standard analysis. I believe this is in part due to the reductionistic tendency of describing collective rationality simply in terms of individually the best choice given the expectation of the other's choice. Then, the aspect of 'collectivity' is redundant, and e.g. the problem of unsatisfactoriness lacks a plausible resolution that emerges from conformity to commonly known social practice, e.g. a social norm or an institution, which applies in a given community.

It is often claimed that the approach that emphasizes and embraces individual rationality is antagonistic to an approach that emphasizes social norms and structures. The antagonism of the two approaches is due to the impression of their basic ideological nature. Game theoretical accounts of social situations are often seen as reductionistic. In contrast, social properties do not reduce to mere beliefs and expectations. Furthermore, the recommendations of rational deliberation and the precepts of social norms are sometimes claimed to be in contradiction with each other. If this impression is correct, then from the point of view of the approach that I am defending, a remarkable part of social reality – viz. social norms, institutions, conventions and customs – would appear to be the product of irrationality. In that case, my starting point that human agents are basically rational beings that conform to social practices in part by virtue of general conformity would be totally misapprehended, and the study at hand would be absurd.

---

anywhere else in this essay, as they would all miss the point.

However, I deny that there is any basic inconsistency between e.g. social habits and rational choices. Occasional conflicts between morals and rationality are better understood from the point of view of accepting that the world we live in is incomplete, and so are our skills and capacities. From this perspective the rules and habits that facilitate social interaction are more than welcome. The real challenge, then, is to fit these apparently antagonistic elements together in order to increase our understanding of the human social reality. For this reason, I find it very attractive to apply a reductive strategy as a part of accounting for social habits that express collective rationality. Social habits can resolve different types of problems of collective action that are often illustrated in terms of game theoretical models, such as the prisoner's dilemma, (**PD**) and the game of pure coordination, (**CC**).[2] However, since I believe that social habits are not describable in ordinary game theoretical terms, the aspiration of applying reductive explanatory strategy requires corrections to standard game theoretical analysis. Hence, I aim at giving an adequate account in terms of a novel class of games – funk games - for I believe that applying reductive explanatory strategies can serve to clarify the otherwise complicated issues. The idea of applying reductive strategies for clarificatory purposes is not new.[3] However, let me stress that I am not committed to ontological reduction of e.g. collective action.

The notion of funk games refers first and foremost to rational interaction situations – games – familiar from e.g. game theory. Within the context of game theory, rationality means approximately the same as maximizing expected utility, and interaction states that rationality is supposed to take place in social settings. In its weakest form this means that in order to make a move that is based on the best possible reasons, an agent must take into account an opponent's interests and deliberations over his own interests and deliberations. Individual rationality emphasizes direct maximization of the expected utility under the given circumstances. However, as already noted, sometimes separate individual deliberations do not suffice as the justified basis for making a satisfactory move. Then, indirect maximization by means of coordination may be needed to achieve this goal in the most effective way. That is, reciprocality of expectations and actions optimizes the chances for success. In addition, when an agent faces a similar situation over and over again, the leap that is necessary for the prevailing practice to be transformed into social habit and, thereby, for acting without present deliberation is very short. If and when such a leap is made, an agent no longer acts *directly* in order to maximize his expected utility, but in order to satisfy the reciprocal expectations under given circumstances. In short, collective rationality has become predominant. But this means that the game no longer refers to a deliberated choice that aims at direct maximization. The maximization of the expected utility is seen

[2] Cf. the next chapter for a detailed discussion of these models.

rather as the function of satisfying the corresponding expectations than as a deliberated choice between the available options. Conformity to a move that accords with mutual expectations describes the behavior of the agents in this situation. The main reason for making the expected move is that it is mutually known to be mutually expected, not that it is the directly expected utility maximizing move, although the function of the conforming move is to maximize the expected utility. This leads to the other constituent of the notion of funk game. The 'funk' in the notion of funk game refers to the functionality of the moves in the given interaction situations. The moves, viz. conformity, realize e.g. social conventions, social norms and social institutions - or social habits. Generally, acting on a norm or following a convention takes place by a routine move that is not presently deliberated. Even so, compliance may be considered to be rational. Funk games are meant to capture the sense of rationality that takes place without present deliberation.[4]

Another obvious point of collective rationality is that sometimes the most effective way to implement one's aims turns out to be impossible without the aid of certain general, commonly accepted routines. Instead of using the sheer force of (ideal) inferential skills and capacity, it is sometimes reasonable to spare that energy e.g. by adopting and conforming to one's or one's community's habits. That is, sometimes the most effective way is to try and adopt a routine that is followed, rather than to stop and deliberate each time over the alternative options, and sometimes the most effective way is implemented by means of coordinated actions, rather than e.g. by trying to maximize one's expected utility directly.

The point of view of this essay, however, sets certain constraints on the rational agents. It makes no sense to hold that ideal agents equipped with ideal inferential skills and perfect capacities have any need to customize their behavior in order to more efficiently reach the best results. In an ideal world, customizing rationality would be redundant. Thus, the point of view is that of ordinary agents equipped with incomplete inferential skills and limited capacities. However, the idea of ideal rationality is not overthrown. It may serve certain clarificatory purposes, especially those revealing the mechanisms of inferences. Furthermore, under certain circumstances the inferences of ordinary agents approximate the ideal rationality. Under such circumstances customs and, consequently, funk games may lose their applicability.

The approach I am defending brings into focus several critical points concerning the standard approaches of collective action and social practices. The picture of rationality that is provided by the standard game theoretical approach turns out to be too narrow. For instance, it does not capture routine

[3] Cf. e.g. Grice (1989), especially Strand Four (pp. 349–359), and the discussion of Meaning.

moves that are made without present deliberation, e.g. conforming to a prevailing practice.[5] All rational behavior does not require deliberation. Furthermore, the machinery of the standard game theoretical approach is unnecessarily heavy for simple everyday situations that involve rational behavior. This is in part due to not being able to deal with the routine moves that are backed by collective reasons, and in part due to operating with requirements for rational behavior that are too demanding to carry out in unforeseen everyday situations. Especially, long chains of backward induction are too demanding for ordinary agents, and they are also often quite redundant for optimal use of one's resources from the point of view of an ordinary agent. There is no point in applying small world rules in the large world that ordinary agents inhabit. It is questionable whether rationality increases remarkably by stretching the chain of recursions of beliefs. In fact, leaning too much on the heavy machinery occasionally leads to resolutions that are counter-intuitive, and sometimes, despite the heavy machinery, the approach is totally incapable of yielding a resolution at all. These observations can be described in terms of the problem of unsatisfactoriness and the problem of indeterminacy, respectively. Finally, the present approach points out that the standard solution concept, viz. equilibrium, does not succeed in delivering a plausible account of social habits, nor does it provide a plausible picture of free riding.

***

The plan of this essay is the following. In Chapter 2, I review the discussion of the problems of collective action and the problem of free riding. The basic idea is that the discussion emerges from two distinct sources that can be traced to the **PD** and the **CC**, depending on the emphasis. In the literature on the problems of collective action the former trait has been predominant. However, this account turns out to be inadequate. For instance, it makes collective rationality epiphenomenal, and it does not succeed in explicating certain basic problems of collective action, e.g. free riding. An eligible expectation of getting a free ride entails a touch of coordination, and so does the expectation of successful collective action. However, the expectations of cooperation and succeeding in coordinating actions seem a bit awkward from the point of view of the direct maximization of expected utility. This is in part because of the requirements of uniqueness and admissibility. These requirements insist that the problems of cooperation are resolved by complying with the unsatisfactory resolution, and that the problem of coordination is reduced to the problem of unsatisfactoriness by allowing randomization of strategies. The alternative characteristic that emphasizes the aspect of coordination in explicating

[4] This means, especially, immediately preceding deliberation, or deliberation that is considered now, in the present tense.

collective action points out that collectively rational action requires interdependence of expectations and actions on the part of the agents. However, the standard game theoretical approach does not offer any means to explicate the interdependence. A mere reference to a social convention does not suffice. In addition, the interdependence of a resolution is not provided just by virtue of its being an equilibrium. Something more concrete is required.

In order to introduce a plausible account, the notion of rationality needs first to be reassessed. This issue is discussed in Chapters 3 and 4. For introducing rationality in terms other than those of direct maximization of the expected utilities, it is essential to take into account that we are dealing with agents that are equipped with limited skills and finite capacity. Especially, in order to justify the importance of collective rationality, the agents under study are supposed to be such that they may benefit from putting their heads together and that they are supposed to adopt certain social practices that facilitate their intercourse with each other. So, in Chapter 3 I lay down the background conditions of 'folk rationality', viz. the rationality of ordinary individual agents. In the core of this view is still the idea of rationality as maximizing one's expected utility. However, the circumstances and the agents are not supposed to be ideal. This means, especially, that even though the hard core of rationality can be described in terms of maximizing expected utility, the folk core faces several constraints that puts an ordinary agent into a position in which maximization is, strictly speaking, impossible. Nevertheless, the corresponding behavior may be argued to be rational. In Chapter 4, I turn into the entailments of the folk view and point out certain welcome properties of it. Most importantly, if a resolution to e.g. a social dilemma appears to be counter-intuitive, it cannot represent the folk view. Moreover, if the resolution is achieved in part by giving up a postulate of mutual belief in rationality, it cannot yield an eligible answer to a social dilemma in terms of the folk view, since the idea of rational cooperative resolution would then be self-destructive. The folk view realizes that despite the imperfect world, mutual expectation of rationality can be enforced, especially in a recurring situation.

The aspect of recurrence or continuance opens up also a possibility of resolving the problems of indeterminacy by coordinating actions. Chapter 5 begins with a discussion of this issue. Conforming to a prevailing pattern is the key to successful coordination. In the literature the prevailing pattern is often expressed in terms of social conventions, norms and institutions. In the current discussion it is often held that social practices can be explicated in broad game theoretical concepts. Especially, social conventions, institutions and norms are thought to be explicable in terms of equilibrium. I show that this account is incorrect. The picture given by equilibrium is too static for the task. In addition, the notion of

[5] Note, however, that 'without present deliberation' does not mean without 'any deliberation'.

choice does not adequately describe the behavior of rational agents when norms or conventions are at stake. Conformity does not reduce to a choice. The difference between the two is easily shown. Choice requires present deliberation over given options, whereas conformity does not. Rational conformity may ultimately be based on deliberation, but the present conforming move may be performed in a routine manner in which no reconsideration of the corresponding present move takes place. Then, the broad game theoretical terms cannot be the standard game theoretical terms. Rather, something else is required. I suggest that the rational interaction that I call the funk game does the job.

The discussion presented in Chapter 5 proceeds as follows. In section 5.2, I present first the initial account originating from David Lewis' (1969) analysis of social conventions. My aim is to provide the preliminaries for the social reasons of collective rationality. However, I realize that the initial account is unnecessarily restricted. It does not deliver all the demanded goods, especially the norms of cooperation. There are two principal ways to approach the issue in terms of rational behavior.[6] In section 5.3, I study an equilibrium account and its options. Three basic options of the organic or non-cooperative approach are studied: (i) giving up the postulate of the common knowledge of rationality, (ii) applying the selective incentives, and (iii) weakening the standard solution concept. None of these candidates succeeds. An alternative option is to give up the equilibrium account and introduce an argument from the self-enforcing path. This is done in section 5.4. Conformity to an existing and mutually known path that provides a resolution to a corresponding social dilemma resolves the problems of the equilibrium account. The path may not strictly provide the condition of equilibrium, and yet conforming to it in part because of the social reasons can be considered to be rational. The norms of cooperation can be explicated within this scheme. However, as already noted, this approach does not take place in standard game theoretical terms, but in terms of conforming to social habits that are expressions of collective rather than individual rationality. Then, consequently, the games under discussion are not standard games, but what I call funk games.

The final chapter of this essay, Chapter 6, begins by pulling together the themes discussed above. I present certain entailments and applications of the approach I am defending with respect to free riding. I argue that the approach that proceeds in terms of ordinary agents and funk games provides a plausible account of free riding. However, the available results are more general than that. A free ride effect can be applied in the analysis of other phenomena of the social world as well. Chapter 6 also points out certain welcome features of collective rationality in general.

[6] These may not be the only ways.

2 THE PROBLEMS OF RATIONAL COLLECTIVE ACTION

According to a widely held view rational action means, approximately, action that maximizes the expected utility. I take this to be the core notion of rationality. Although the characterization is vague and even controversial, it will do for the present purposes.[7] Rational collective action, in turn, aims at maximizing expected utility under constraints of realizing mutual benefits. Collective rationality cannot take place unless coincidence of interests predominates. Collective rationality is, however, an even more controversial notion than individual rationality is. It is agreed that collective rationality arises from individual rationality, but occasionally it contradicts with this, and sometimes the collective rationality corrects the individual rationality. That is, by means of collective rationality people can effectively maximize their expected utilities that otherwise would be virtually impossible. Despite the difficulties that seem to beset this argument, I hope the present essay will clarify this. An initial step in this direction is to clarify the problems of collective action and their relation to rationality.

The problems of rational collective action are thought basically to emerge from two distinct sources: problems in coordination of actions and problems in cooperation. The former represent problems of indeterminacy and the latter represent problems of unsatisfactoriness. The most common way of illustrating these two types of problems has been in terms of the game of pure coordination, (**CC**), and the well-known problem of the prisoner's dilemma, (**PD**). The reasons for seeing the problems of collective action in terms of these polar cases have their roots in the classical literature concerning the social contract theory and the emergence of society.

Those who have emphasized that rational collective action is a matter of successful coordination have held that there is no basic conflict that requires resolution before a society may arise. They, together with e.g. Jean-Jacques Rousseau, have thought that social institutions are primarily sets of conventions that have emerged in the course of history. This view also emphasizes the corrective aspect of collective rationality. Many problems become resolved simply by conforming to social habits.

According to this view, individual deliberations cannot guarantee the beneficial effects of conforming to the prevailing customs. In contrast, the proponents of the conflict view have held, in accordance with e.g. Thomas Hobbes, that because of the antagonistic interests of human beings, no society is possible without an external authority. Furthermore, collective rationality has no effect unless the conflict is resolved. According to this view, the implementation of collective rationality is determined solely by the collection of individual interests, as in the invisible hand explanation.

The present study presupposes that, despite of their distinct sources, the two accounts of collective action complement each other. Especially, no authority is necessary for the emergence of cooperation even if the interests are in conflict. And, furthermore, many social dilemmas that otherwise would be left unaccounted for become explicable in terms of an analysis that is sensitive to the properties of conflict and coordination in coexistence. For instance, the free rider problem becomes naturally explicable when the relevant details of both sources are taken into account, which is not true of an analysis that goes either in terms of conflict or in terms of coordination only.[8]

The study of the distinct sources in conjunction has also a very important advantage as opposed to treating the two approaches in separation. Namely, indeterminacy and unsatisfactoriness apply to the very same game theoretical models, depending on which features are emphasized, and often problems belonging to one source are resolved by transforming them into a case belonging to another. For instance, the problem of unsatisfactoriness is resolved by applying a more dynamic account and relaxing the postulate of uniqueness, thereby making satisfactory outcomes rationally conceivable. Consequently, due to introducing alternative solutions in addition to the existing ones, the problem of indeterminacy becomes inherent. Furthermore, to avoid the problem of indeterminacy, some authors have suggested that mixed strategies provide a unique solution to the problem of deliberating the correct choice.[9] Consequently, the solution to the problem at hand means introduction of the problem of unsatisfactoriness, and consequently the problem of collective action awaits resolution. Avoiding this apparent circle requires an elaborate study of the problems of collective action and their relation to each other. Otherwise one of the problems is solved by introducing the other. The approach that considers the sources of the problems in coexistence makes it possible to study the problems of collective action as a whole and to show that the emergence and the resolution of the problems has to do with their

[7] Rationality and rational behavior will be discussed in more detail in Chapter 3: 'The Folk Core of Rationality'. For a detailed discussion of rationality, cf. Rescher (1988).
[8] The corresponding accounts that I find inadequate are suggested e.g. by Pettit (1986): **PD**, Taylor (1987): **CG**, and Hampton (1987): **BS**. All of them try to account for the free rider problem in terms of a single source of conflict or of coordination.
[9] Cf. e.g. the case of the **CG**, or Harsanyi's (1977) suggestion about **PD**-type problems.

relation to each other. The basic problem then is to construct a model that takes the distinct elements into account in a way that collective action and the problems it may face are rationally conceivable.

In order to motivate the study of rational collective action I need first to study the problems of collective action and their basic sources in separation, since studies in terms of the sources of problems in conjunction are rare. Since the current discussion of the problem of collective action has mainly emphasized the points of view of conflict, it seems natural to begin the discussion with this approach, as well. [10]

2.1 I Know What's Best for Us but I'm Not That Stupid to Do It

The presentations of the problems of collective action usually begin with observing an example that is said to be a case of the Prisoner's Dilemma (**PD**). A basic, single-shot **PD** is generally characterized by two distinct properties that can be easily observed in Figure 2.1.1.

	C	**D**
C	(3, 3)	(1, 4)
D	(4, 1)	(2, 2)

Figure 2.1.1 the Prisoner's Dilemma

These two properties are (i) **D** is the dominating *strategy* for each player, and (ii) the *outcome* **CC** jointly dominates the equilibrium outcome **DD**. From these properties it follows that in the case of a single-shot **PD** the rational agents have no other option than to choose the dominating strategy. So, irrespective of the opponent's choice strategy, **D** maximizes the expected utility. That is, choosing **D** is the rational option and the only rational option for an agent. However, since the equilibrium outcome that results from both agents choosing the dominating option **D** is jointly dominated by the non-equilibrium outcome of both choosing their individual dominated option, the game leads to a collectively unsatisfactory result. So, the individually dominant option and the collectively dominant option clash. Intuitively, both agents know what's best for them as a group or collective, but neither is stupid enough to make a corresponding move toward that outcome. This is the basic motivation for the emergence of the problem of collective action.

[10] Cf. Tuomela (1992). He asserts (p. 172) that the problems of coordination are not at stake when discussing the problems of collective action. Cf. also Hardin (1982), Taylor (1987) etc. For an opposing view, cf. Hampton (1987), who emphasizes the aspects of coordination in explaining the free rider problem.

2.1.1 The Basic Example

Let us apply this motivation to an example of the problem with the public goods provision.[11] The discussion of problems of collective action proceeds generally accordingly. We have spotted a model that emphasizes a serious problem in achieving a mutually accepted outcome, and we know that due to their general properties, public goods and their provision are vulnerable in the face of corresponding deliberation. The classic approach holds that the **PD** is the model that provides a solid explanation for the emergence of the problem of collective action.[12] So, let us suppose for the sake of argument that whenever it is a matter of providing a public good (viz. a non-excludable and jointly supplied good) for a community, the community faces a **PD**-type problem. Russell Hardin (1971, 1982) illustrates the clash between individual and collective best choices by providing an example of a group of ten members whose common interest is to provide a public good whose value twice its cost. [13] Especially, it is supposed that if all members pay 1 unit, the benefit to each will be 2 units. When the costs are subtracted from the benefits, in case of all paying, the outcome will be 1 unit to each member. Hardin's calculations of the payoffs in a game of individual vs. collective (1982, p. 26) amounts to the matrix shown in Figure 2.1.2.

		Collective	
		Pay	Not pay
Individual	Pay	(1, 1)	(-0.8, 0.2)
	Not pay	(1.8, 0.8)	(0, 0)

Figure 2.1.2 Individual vs. Collective

From the point of view of Individual, 'Not pay' is the dominating strategy. Since it is individuals who make the choice whether to pay or not, Collective's choice will ultimately, according to Hardin, be

[11] Viz. a non-excludable and jointly supplied good. Cf. e.g. Taylor (1976) and Hardin (1971) for discussion.
[12] Cf. Olson (1965), and Hardin (1971).
[13] Hardin speaks of a collective good instead of a public good.

whatever Individual's choice of a strategy is. Under the circumstances of dominance no one pays. The payoff will be 0 unit in the given example. Thus, the example represents a **PD**-type situation.

Of course, Hardin's example is unnecessarily restricting for general purposes, but it succeeds in illustrating the clash of collective vs. individual. Although Hardin's example has been criticized for being incorrect and misleading, it contains certain welcome properties for explicating *the free rider problem*. [14] Most importantly, a relevant point is stated in presenting the case in terms of a game between individual and collective, even though the way Hardin states the case is oversimplified. Especially, the notion of rationality and the essential properties of rational agents in social interaction need elaboration. In any case, among the insights that can be gained from the manner in which the example is presented is that in free riding the provision of the collective good seems self-evident. The game of Individual vs. Collective makes it self-evident from Individual's point of view, since paying is the dominating strategy for Collective. That is, on the collective level, public goods are rationally conceivable. However, the problem in accounting for the motives illustrated in the game is that under conditions of ideal rationality no one can expect a cooperative move in the given situation. This is a problem that I hope to solve in this essay.

2.1.2 A Generalization of the PD-type Problems

A remarkable feature of Hardin's example is that it emphasizes how the individual benefit of not paying can be enormous compared with the collective loss. That is, a single free rider sees his move as virtually harmless from the point of view of the group as a whole, while individually the profit is remarkable. It is also important to observe that the **PD**-type problems can be constructed in a variety of ways. Philip Pettit (1986) tries to capture this feature by introducing A-type **PD**'s. An A-type **PD** is such that in a many-party dilemma no cooperator is made worse off by a lone defector than he would be under universal defection. This seems to be exactly the case in the game of Individual vs. Collective. In contrast, a B-type **PD** represents a generalized standard **PD**, since then a lone defection plunges some cooperator or cooperators below the baseline of universal defection.[15]

Pettit's distinction between type A and type B problems points to a more general characterization of the problem that e.g. a standard **PD** represents (see Figure 2.1.1). A more general characterization of **PD**-type problems is, in fact, supported in the literature. For instance, John Harsanyi (1977, 280) has suggested that any problem in which a rational agent must choose a course of action *S*

[14] Cf. the discussion in Hampton (1987) and in Tuomela (1992). Cf. also Taylor and Ward (1982).

even though a course of action *S** would yield a better result than the course of action *S* for all is a **PD**-type problem. There may be several reasons this: (i) *S** is not in equilibrium; (ii) *S** is in equilibrium, but it is not stable enough; (iii) there is a bargaining deadlock; or (iv) there is a coordination deadlock. Thus, Harsanyi argues that besides the standard **PD**, the games illustrated in Figure 2.1.3 all present the **PD**-type problem.

I:	**B1**	**B2**
A1	(2,1)	(0,0)
A2	(0,0)	(1,2)

II:	**B1**	**B2**
A1	(0,0)	(1,1)
A2	(1,1)	(0,0)

III:	**B1**	**B2**
A1	(2,2)	(1,2)
A2	(2,1)	(1,1)

Figure 2.1.3 Variety of games representing a **PD**-type problem

These games represent the bargaining problem, the coordination problem and the indifference problem, respectively. Game I has three equilibria, of which two jointly dominate the third one in the strict sense. The third equilibrium is unprofitable for both players, even though it is the maximin point of the game. The equilibria are *S** = (A1, B1), *S*** = (A2, B2), and *S* = (A1+A2, B1+B2). Game II also has three equilibria. The solution of Game II corresponds to the solution of Game I. In other words; the individually best choice provides an outcome which is jointly dominated by the collectively rational outcomes. In Game III all possible pairs of choices are equilibria. *S** = (A1, B1) jointly dominates the other pair of choices in the weak sense. However, when played strictly noncooperatively, *S** = (A1, B1) is unstable because of the indifference problem. The only stable, and thus applicable, equilibrium is *S* = (A1+A2, B1+B2). In short, acting according to the rule of the best reply and with no other information about the situation yields an unsatisfactory result for all.

If there were a way to coordinate their actions, rational agents would do so in the games illustrated in Figure 2.1.3, since suitable coordination would provide a collectively dominating outcome over other alternatives. In the absence of a resolution that guarantees a coordinated solution, the agents have no other possibility than to try to seek for a unique expected utility-maximizing option. As such, the models in Figure 2.1.3 represent the problem of multiple equilibrium, which, in turn, may make the resolution of the game indeterminate.[16] In the search of a unique solution to the situations representing the problem of indeterminacy, randomizing has proved useful. Then, an individual's utility maximizing strategy turns out to be collectively dominated by the coordinated strategies in the models under

[15] For a more detailed discussion cf. Pettit (1986).

discussion. The normal form of presentation of these models does not give any clue for implementing the coordinated outcomes. Consequently, the requirement of uniqueness of the solution on its part leads to a **PD**-type problem. So, in a sense, the problem due to multiple equilibrium is resolved by putting an emphasis on the uniqueness, and, consequently, transforming the problem into that of *unsatisfactoriness*. As a result, the common denominator of the standard **PD** game and games I, II, and III is that the outcome of the game will be suboptimal, given that the game is played noncooperatively and the players are rational. Generally, in all games representing a **PD**-type problem, the solution is *unsatisfactory* for all. In short, the generalization of the **PD**-type problems is realized by reducing the problems of collective action to the problem of unsatisfactoriness. That is, the problem of unsatisfactoriness illustrated by the **PD**-type problems is one of the main sources of the problems of rational collective action. Individual rationality hinders collective aspirations also when the **PD** is, strictly speaking, not involved.

2.1.3 Problems of Collective Action as the Product of Unsatisfactoriness

Jon Elster (1985) distinguishes between a strong definition of the problem of collective action, which he identifies with the **PD**, and a weak definition of the problem, which is meant to capture non-**PD** problems, as well. The weak definition requires that (i) the cooperative solution jointly dominates the universally noncooperative solution (or universal defection from the possible cooperative solution), and (ii) the cooperative solution is individually unstable *and* individually inaccessible. According to Elster, individual instability prevails if each individual has an incentive to defect from universal cooperation, and individual inaccessibility prevails if no individual has an incentive to take a first step out of universal noncooperation. Together, these conditions amount to an implementation of the problem of unsatisfactoriness.

However, Elster points out that even the weak definition of the problem of collective action is inadequate, since there are cases of the problem in which cooperation is either individually unstable or individually inaccessible, but not both. This means that the problems of collective action cannot be explicated in terms of unsatisfactoriness only. Despite this fact, it can be said that the problems that are described in terms of instability and inaccessibility in a sense relate to the problem of unsatisfactoriness. Alone either of them provides a partial problem of unsatisfactoriness, and all together makes for a full problem, so to speak. Although the solid base of unsatisfactoriness as the source of problems is

[16] Cf. Harsanyi & Selten (1988) for a detailed discussion of the sources of the problems of rationality.

weakened by this way, Elster's definition can be easily corrected by focusing on the second clause of the weak definition. That is, (ii) the cooperative solution is individually unstable *or* individually inaccessible *or* both. Elster mentions Chicken (**CG**) and Assurance (**AG**) as examples of the cases concerned with either inaccessibility, or instability, but not both. The two-person matrices of these games are shown in Figure 2.1.4.

	Chicken			*Assurance*	
	C	**D**		**C**	**D**
C	(3,3)	(2,4)	**C**	(4,4)	(1,2)
D	(4,2)	(1,1)	**D**	(2,1)	(3,3)

Figure 2.1.4 Chicken and Assurance

The main point of Elster's definition of the problem of collective action is that the problem has to do with the issue of unsatisfactoriness that is due to the clash between individual and collective rationality. That is, the outcome will be unsatisfactory, since the implementation of collective rationality is either inaccessible or unstable or both. Thus, even if the situation cannot be described strictly as a case that provides the problem of unsatisfactoriness, the implementation of collective rationality is obstructed because it is individually inaccessible or unstable, and the outcome is, consequently, unsatisfactory from the collective's point of view. Now, when the problem of collective action is seen from the point of view of unsatisfactoriness owing to inaccessibility and instability, it is possible to distinguish between the problems that represent pure unsatisfactoriness and the problems that represent incomplete unsatisfactoriness. The main class consists of **PD**-type problems, and the minor classes consist of those of the type **CG** and **AG**. Although the **AG** hardly represents a true problem of collective action,[17] it turns out to be a crucial model for seeking the resolution to the problem of unsatisfactoriness.[18]

It is worth noticing also that Elster's definition of individual inaccessibility can be emphasized in two alternative ways. Individual inaccessibility prevails if either (a) *no one* has an incentive to *X*, or (b) everyone has an incentive *not* to *X*. *X* means naturally the act of taking the first step from universal noncooperation. The first alternative suggests a weaker case of inaccessibility: an agent may have an incentive not to *X* or she may just be indifferent between *X* and not-*X*. In other words, an agent weakly prefers not-*X* to *X*. The second alternative may, correspondingly, be called 'strong individual

[17] Cf. e.g. Taylor and Ward (1982), and Tuomela (1992) for discussion.

inaccessibility' in contrast to the weak case, stating the strong preference of not-*X* to *X*. Which one of these two alternative interpretations Elster applies does not show up in the context.

The notion of individual instability can be given two alternative emphases as well. According to Elster, individual instability prevails if every one has an incentive to *Y*, where *Y* is taken to mean the act of taking the first step from universal cooperation. In that case, an agent strongly prefers *Y* to not-*Y*. According to Elster, then, the outcome is unstable because it is not in equilibrium. However, an outcome may be unstable even if it *is* in equilibrium. The instability problem may emerge for instance in the case of indifference. It may happen that all the outcomes are in equilibrium, thus causing the solution to be unstable. Consequently, the weak definition of individual instability in contrast to the strong definition can be formulated as follows: an agent weakly prefers *Y* to not-*Y*. In other words, either an agent has an incentive to *Y*, or she is indifferent between *Y* and not-*Y*. Thus, I suggest that for an agent to have an incentive to choose between this or that course of action must be taken to mean that the agent is not indifferent between the possible outcomes to which the correspondent actions are directed.

Now, on grounds of the satisfaction of inaccessibility and instability it is possible to classify the problems of collective action which emerge from unsatisfactoriness. The common denominator of the problems discussed here is that cooperation is beneficial from the point of view of the corresponding group. In its strongest form the weak definition states that (1) each individual derives greater benefits under conditions of universal cooperation than he does under conditions of universal noncooperation, and (2) cooperation is individually unstable and inaccessible. Furthermore, if an individual has an incentive to defect from the possible universal cooperation (individual instability), then it is not in equilibrium. In addition, if an individual has an incentive not to take the first step away from the possible universal noncooperation (individual inaccessibility), then that outcome is in equilibrium. When both parts of the second clause are in force and the agents are not indifferent to which alternative to choose, each individual derives more benefits if he abstains from cooperation, regardless of what others do. In other words, strong instability and strong inaccessibility mutually generate the strong dominance of the noncooperative strategy when the two-person case is involved. As this situation is equivalent to Elster's strong definition of collective action problem, it is a standard case of the **PD**.

When the notions of instability and inaccessibility are applied in their weak sense, the scope of applicable models is increased. The problem is of the **PD**-type, since the only case for which it holds that (i) universal cooperation is preferred to universal noncooperation by every individual, and (ii)

[18] Cf. Chapter 5 for the assurance problem.

universal cooperation is individually inaccessible or unstable, *and* (iii) universal noncooperation *is* in equilibrium, is a case of the **PD**-type problem. Note that in its weakest form, the definition does not rule out even the possibility that the cooperative outcome is in equilibrium, since individual inaccessibility and instability do not rule out equilibrium. Thus, universal cooperation may be an equilibrium outcome at the same time as it is strictly preferred to universal noncooperation by everyone; however, the **PD**-type problem still emerges, since the players do not have any way to figure out how to coordinate their actions. Furthermore, when it is accepted that the problems of collective action may arise even when either instability or inaccessibility but not both are in force, also the **AG**-type and the **CG**-type problems become available for the discussion that is concerned with the problem of unsatisfactoriness.

However, the approach that reduces the problems of collective action into problems of unsatisfactoriness runs into troubles almost immediately. The main reason for this is that collective rationality is treated as an *epiphenomenon*. Collective rationality works only as a counter-factual test of expectations and hopes of rational individuals. It is of no real use, since the direct maximization of expected utility by individuals is what counts, and in the given examples there is no such thing that can be genuinely considered as collective rationality, that is, as maximization that would be implemented by means of suitably coordinating actions. Instead, there seems to be only a problem in which a collection of rational individuals does not seem to reach an intuitively acceptable and satisfactory outcome. To see this, the study of free riding is instructive.

2.1.4 Free Riding: The First Approximation

The problem of unsatisfactoriness is often considered as a symptom of free riding. In the language of economists, when the provision of public or collective goods takes place at a suboptimal level, there are good reasons to suspect that some agents are benefiting from the others' efforts without making a contribution of their own. Especially, some individuals find it rational to take advantage of the willingness of others to contribute to the collective good in a way that threatens its production. Moreover, specifically, those individuals may do this despite the risk that there will not be enough contributors, since that is the way for them to maximize their expected utilities.[19]

In order to show that the explanation that goes in terms of the problem of unsatisfactoriness is adequate, certain requirements must be met. Especially, it must be shown that individually the best choice leads to a suboptimal outcome. This is, naturally, true of **PD**-type problems in general, and if it is

[19] Cf. especially Pettit (1986), Hampton (1987), and Tuomela (1992) for discussion of the free rider problem.

true of free riding, then unsatisfactoriness may account for the emergence of free riding as well. This is the line of argument that e.g. Philip Pettit (1986) is committed to.

Pettit presents four conditions under which a *paradigm free rider problem* arises. He claims that whenever a "free rider problem" meets these conditions, it is a case of the A-type **PD**, that is, a version of those **PD**-type problems in which a lone defector does not harm the cooperators in the sense that any agent is forced under the baseline of universal defection. Since Pettit claims that the problem is a paradigm case, it must represent the most typical case of free riding. However, I shall show that, although valid at least as far as a **PD**-type problem is concerned, Pettit's analysis is not adequate to account for free riding under the described circumstances.

According to Pettit (1986) the paradigm problem arises under four conditions.

> 1. There is a nonexcludable good attainable for a group: that is, a good enjoyed by all, though perhaps with different intensities, if enjoyed by any.
> 2. It can be attained, and rationally attained - whether at one of many levels, however smoothly related, or at only level possible - by K members of a group, where K is less than all members. The reward to each contributor in a subgroup K will exceed the cost of his contribution, though not necessarily by the same margin.
> 3. It cannot be attained, or at least rationally attained, by just one member of the group: the reward to the lone contributor will not cover the cost he has to bear.
> 4. The fear of contributing when the good is not produced, and the hope of not contributing when it is, make it rational for each person not to contribute to the production of the good: specifically it means that the strategy of not contributing maximizes expected utility. (1986, pp. 367-8)

Pettit's analysis has been criticized on the grounds that the A-type or the paradigm free rider problem fails to meet the conditions of the **PD**, especially dominance (and also suboptimality).[20] I imagine the criticism is mainly due to indeterminacy that generalizing the **PD** creates.[21] Upon closer examination, however, this criticism is unfounded, and Pettit's analysis turns out to be valid. Especially, in case of the A-type problem from the individual's point of view, the condition of dominance holds, and when taking for granted that the paradigm problem converges to the A-type, the dominance holds for the paradigm problem as well.

Since the generalization presented by Pettit leaves open how an arbitrary agent should rank an outcome in which K – 1 agents cooperate, there appears to be incontinuity in dominance. That is, it appears that in case of K – 1 contributing, anyone would prefer to contribute. Note that I am not saying anything about the agent knowing whether there are exactly K – 1 other agents contributing, but about an agent's attitude towards there being such a possibility. The argument from incontinuity of dominance may be due to seeing the case as a game of Individual vs. Collective (see Figure 2.1.2). The situation

[20] Cf. e.g. Tuomela (1988). Cf. also Hampton (1987) for detailed discussion on free riding and the **PD**. It is worth notifying that Pettit himself also points out situations in which the dominance does not hold: that is, when the good is lumpy at level K.
[21] Cf. Pettit (1986), p. 364 for an argument for the generalization.

may be such that there may be a point of incontinuity of dominance from Collective's point of view (that is, level K). That is, there is a point (level K) at which Collective is still above the baseline of universal defection. As a result, it is mistakenly argued that Individual would think from Collective's point of view and maximize his expected utility by cooperating at that point. Then, the condition of dominance is not met, and the situation fails to represent the **PD**. Or, at least, the expected utility maximizing choice does not lead to an unsatisfactory result, in which case the situation fails to represent any **PD**-type situation. However, under the given description, Individual makes his deliberations from his own point of view, in which defection dominates, not from Collective's point of view, in which it does not dominate. As a result, Individual has no other option than to maximize the expected utility by defecting.

There is another possibility of dealing with the incontinuity argument. It can be and, in fact, is also argued that when a collective good is *lumpy* at level K, the circumstances of providing the good fail to represent a **PD**-type problem.[22] This being the case, when an agent *knows* that K – 1, and only K – 1, others have contributed or are going to contribute, it is argued that under these circumstances anyone will prefer to cooperate. Hence, there *is* incontinuity in the dominance. This argument, however, fails to be sound what comes to refuting cases involving **PD**-type problems and the motives for free riding, for it requires knowledge on the part of the agent that cannot generally be required, and it fails to represent a situation in which rational behavior on part of the agents would lead to an unsatisfactory outcome. Furthermore, the argument shows the opposite of the motivation for free riding quire well. Namely, it is generally accepted that free riders maximize their utility by defecting, even if this would endanger providing for the collective good. So, even if the collective good in question were lumpy at level K, we must be dealing with a **PD**-type situation, if not a standard **PD**, since free riding has something to do with the problem of unsatisfactoriness, that is, with the conflict between individual and collective rationality. Thus, it suffices to say that despite the incontinuity of dominance in the case of lumpy goods, an agent maximizes his expected utility by not contributing, and following this expected utility-maximizing strategy leads to an unsatisfactory outcome – be it non-provision of the good or provision at a suboptimal level. This result is due to the given description of the problem. Under a different description of the situation, e.g. in the case of sequential games, other options may come available. Then, contributing may be the only admissible choice and free riding an irrational option. Under the given description, however, this consequence is not warranted.

[22] Taylor and Ward (1982) suggest that this is an instance of **CG**. Hampton (1987) argues for **BS**. Cf. also Tuomela (1982) for discussing various alternatives.

Still, there seems to be something wrong with the argument at hand. Undeniably, free riding must be explicated from the point of view of unsatisfactoriness, but it cannot be adequately understood if the cooperative option is not rationally conceivable as well. Even if the argument from unsatisfactoriness were valid, as it in fact is, and we accepted that free riding on collective action wells up in **PD**-type situations, the analysis of free riding is still inadequate. When seeing the case from the point of view of a rational individual, the apparent dominance (or at least the unsatisfactoriness due to following the expected utility maximizing strategy) leads to failure to contribute and ultimately to non-provision of the good. If rational, then no one can expect to get a free ride under described circumstances. Thus, the model requires elaboration.

The inadequacy of the paradigm model gives a sufficient reason for trying to seek the answer elsewhere. The basic insight is that in order to understand free riding, the element of coordination in collective action should somehow be taken into account in the analysis. This is the first crack in the argument from unsatisfactoriness alone. It has created grounds for reconsidering the problems of collective action.

2.2 A Touch of Coordination

Since the explication of free riding runs into trouble when studied only in terms of unsatisfactory outcomes that arise from the conflict of individual and collective rationality, some authors have tried to approach the issue from the point of view of the elements of coordination.[23] The discussion has been taken place by introducing and defending certain game theoretic structures other than **PD**'s without giving a full analysis of the problem. The problem with these alternative approaches is that although they are for obvious reasons uncomfortable with reducing the problems to unsatisfactoriness, they either implicitly do so or otherwise do not succeed in delivering an adequate explanation. Still, despite their shortcomings, the accounts that emphasize the aspects of coordination seek the solution in the right direction.

2.2.1 Draining a Meadow: a Classic Example

A classic example of the free rider problem connected with the problems of coordination was given by David Hume:

[23] Especially, Taylor & Ward (1982), Hampton (1987) and Tuomela (1992).

Two neighbours may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each other's mind; and each must perceive, that the immediate consequence of his failing in his part, is, the abandoning of the whole project. But 'tis very difficult, and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expense, and wou'd lay the whole burden on others. (III, ii, vii [1978, 538])

Hume's example states clearly the aspect of coordination in the two-person case. It is clear that if either of the neighbors fails to cooperate, the whole project fails. That is, the goals of the agents are interdependent.[24] The generalization of the problem makes the situation less clear. Hume seems to suppose that in case of a relatively large group an occasional free ride does not put the collective goal at risk of not being reached at all. That is, a singular defection does not force any one under the baseline of universal non-contribution, and the defector may enjoy the fruits of others' efforts. This is exactly what happens in the A-type **PD**, but because it emphasizes the element of dominance, the A-type **PD** is considered suspect. Furthermore, the same suspiciousness is thought to hold for the **PD**-type problems in general, since the individual rationality outweighs the collective rationality. Hence, the aspects of coordination in the two-person case of Hume's example are emphasized in the generalizations as well. In order to emphasize the aspects of coordination belonging to free riding, the problem in the standard approach is seen as involving the achievement of equilibrium. From the equilibrium theoretical point of view the models of **CG** and **BS** have been of special interest, depending on the emphasis.

2.2.2 Hampton's Suggestion

Let us begin the study of the equilibrium theoretical approach from the coordination side of the game theoretical spectrum.[25] Hampton (p. 246) quite correctly claims that many collective action problems are questions of coordination rather than conflict dilemmas. Especially, she claims that many situations that are exposed to free riding do not require external authority for sanctioning. She bases her argument essentially on the assumption that the structure of the situation is such that the provision of a collective good is rationally conceivable from the individual point of view even if some agents were illicitly benefitting from efforts of others. So, free riding must take place under circumstances of equilibrium. Hampton suggests that the game of **BS** is the closest candidate for being the model of free riding. This suggestion is backed by an argument that Hampton bases on an analysis of Hume's initial example:

1. Draining the meadow is a collective good: i.e., it is indivisible, non-excludable, and a benefit to the group.

[24] Cf. Schelling (1960) for an introduction to the games of interdependent solutions, i.e. games of coordination.
[25] Cf. Hampton (1987), p. 253.

2. It is a step good. It does not make sense to say that the drained meadow can be "incrementally increased" in either quantity or quality after it comes into existence.
3. Individual production costs and benefits from the good are well defined and commonly known, so that individual preferences for producing the good are commonly known.
4. The group involved in producing the good is what Olson (1965, pp. 22-36 and 48-50) calls "latent" as opposed to "privileged" because there is no individual in the group for whom V_i-C_T>0. Here V_i is the amount of benefits to the *ith* individual and C_T is the total cost.
5. Production costs can be split in a variety of ways among the 1000 group members (that is, the group can define production units in a variety of ways and assign any number of group members to these units). But the minimum number of people capable of producing the good is two.
6. Finally, individual costs to produce the good are not retrievable. An individual cannot recoup whatever he pays to drain the meadow (e.g. monetary costs) before the good's production is completed. (1987, 251)

Because of reasons of facilitation Hampton studies a three-person case. In the example presented by Hume the dynamics in a many-party case are relevantly different from the case of pure coordination that the two-person case represents. The idea of the different dynamics can, however, be captured by a three-person example. Accordingly, it is unnecessary to commit oneself to an example in which one must study the different options that an individual might have in a case of 1000 persons. Especially, three-person cases can be applied to show the different options of one party defecting while the two others are contributing to the production of a collective good in case that the good is lumpy at level K = 2, i.e. in a situation in which some are coordinating their actions while there is a possibility to have a free ride as well. It is only under circumstances in which an agent can rationally expect contribution on part of the others that an agent can expect to get a free ride, and according to Hampton this is the case when coordination of action predominates. However, since getting a free ride is clearly the most preferred outcome on the part of an individual, there is an element of conflict in the coordination situation at hand. The closest available example that fits this description is, according to Hampton, the **BS**.

Now, in order to emphasize the elements of coordination in the given example, the interdependent character of the outcome should be taken into account. Hampton's answer to the problem is that the free rider problems that fit the given description are involved in the *selection* of contributors. That is, supposing that we are dealing with an example of three persons, the problem of the corresponding agents is to reach an agreement on who are the two contributors, and who is the one to hit the jackpot and benefit the efforts of others. This, in turn, means that free rider problems must be involved with the coordination equilibria that can be reached by means of an explicit agreement or by means of conforming to a prevailing convention. This is a crucial point for understanding the nature of free riding, although Hampton's analysis as such must be considered incomplete in many respects. The point, however, is that the expectation of coordination and the realization of coordination equilibrium

mean that providing for the collective good can be rationally expected. This in turn means an opportunity to free ride.

If we suppose for the sake of argument that Hampton's example is sound, then the problem of selecting the contributors represents a three-person *bargaining* deadlock. Under the given conditions there is no way that the three persons can solve the problem of selecting the contributors. There are three coordination equilibria that are equally strong, and each person prefers a different equilibrium than the others, and, furthermore, the fact that there are several equilibria that are equally forceful (from the collective's point of view, at least) creates a problem of coordination. That is, the situation represents the problem of indeterminacy: there is no *a priori* way to determine the rational course of action, since the option of transforming the problem of indeterminacy into the problem of unsatisfactoriness is not available if the coordination of actions is at stake. Before the problem of indeterminacy is resolved, no one can form any expectations about each other's actions. Hence, Hampton's approach is not adequate for explicating free riding, either.

Let us, however, for the moment suppose that the agents have found a way to resolve the problem of determining the rational course of action in the described situation. Suppose, for instance, that in case of selecting the contributors, the agents simply toss a coin or apply another similar simple mechanism to resolve the bargaining deadlocks. (It really does not matter if tossing a coin creates higher-order problems. The point is only to show what Hampton's analysis would entail in case the deadlock were resolved.) Then due to the preference rankings the agents would coordinate their actions in such a way that one would be free riding and the two others would be contributing to the production of the collective good. Neither of the two contributors would have an incentive to defect, since the corresponding outcome satisfies the equilibrium condition. Consequently, the situation is not a **PD**, since the element of dominance is missing; nor is it a **PD**-type situation, since the best reply to the others' moves depends on the position of the agent. That is, each agent wants to avoid the job, but each agent also wants the job to be done rather than see the project abandoned.[26] Since, according to Hampton, coordination of interests predominates in the given example, and since there is also an element of conflict, the situation presents a game of **BS**. I agree with Hampton that the **BS** and the other problems of coordination constitute a serious problem of collective action, but I think it fails to provide an accurate description of free riding. I do not feel comfortable with the idea that a group of people would agree on a project or adopt a habit in which some of them enjoy a free ride while the others are doing all the work, and that this would be the most effective way of coordinating actions.

[26] Cf. Hampton (1987), p. 252.

Described in this way the case illustrates an example of the division of labor rather than a situation in which an agent who should take part into the efforts defects instead. The element of "illegality" is missing from Hampton's suggestion. Instead of defecting the one who gets the free ride is coordinating with others. Of course, it is a matter of interpretation whether the third person in the **BS** represented as in Hampton (1987) can be called a free rider. In my opinion, free riding in general takes place when one defect from the mutually accepted course of action, not by coordinating one's action with the others' actions in a way that can be considered as acceptable.

2.2.3 The Element of Conflict

In the standard accounts of free riding the emphasis is on the conflict of interest, although certain aspects of coordination must be inherent. The idea is that in providing for a collective good the persons involved are expected to contribute, but they do not. I mean that these persons fail to coordinate their actions in a suitable way. However, since it is assumed that the option of free riding is the most preferred one, the failure of coordination is not a catastrophe from the point of view of the free riding person. On the other hand, since the equilibrium theoretical approach holds that in order for free riding to be a rationally conceivable option, the provision of the collective good must also be rationally conceivable from an individual's point of view, such that from the point of view of those who succeed in coordinating their actions the failure of some others is not a catastrophe, either. That is, no one seems to mind if someone fails to cooperate, although the contributors hope that no-one fails. Hence, free riding cannot be described strictly in terms of the coordination equilibria.

An example in which the dominance does not predominate, and in which the failure of "coordination equilibrium" does not lead into a catastrophe, is the game of **CG**.[27] In this game the element of conflict predominates, so that individually the best outcome, viz. free riding, clashes with the collectively most acceptable outcome. The dynamics of the game are such that each agent has a clear incentive to be the first one to defect from the collectively rational strategy. But because the case of all defecting leads to a catastrophe also from an individual's point of view, some may think it too risky to defect, while the others who are daring enough may think that because of these dynamics one could gain from the others' aversion of risk. The game of **CG** (Figure 2.2.1) seems to distinguish between two types of agents that may coexist so that the condition of equilibrium is satisfied. In other words, the daring ones free ride at the expense of the risk-averse ones.

	C	**D**
C	(3, 3)	(2, 4)
D	(4, 2)	(1, 1)

Figure 2.2.1 Chicken

The matrix depicted in Figure 2.2.1 shows clearly that the risk-minimizing and the profit-maximizing strategies clash, and that those who are willing to take a risk profit at the expense of those who want to play it safe. The dynamics of the **CG** fits to the characterization of free riding perfectly. Free riding puts the provision of the collective good at risk, but a free rider will not reconsider this possibility, since this is the way he maximizes his expected utility. This feature is strengthened due to the fact that in the game of Chicken *precommitment* plays a crucial role.[28] That is, it pays to commit oneself to the strategy of defection, thereby forcing the others to contribute to the collective goal. Undoubtedly, this is what occasionally happens in collective action. Thus, the **CG** superficially seems to provide a very plausible approximation for the emergence of free riding.

However, under the given description, a distinction between the different types of agents is not warranted. The division of the agents into two classes in order to explain free riding is methodologically and metaphysically dubious. The distinction between the types of agents is external to the game theoretical model at hand. Nothing in the definition of the situation tells who belongs to which class and how the members of the different classes can be recognized. If the players are treated *a priori* as equally strong, and there are two eligible outcomes, the solution will be indeterminate. The problem is shifted to the higher-order deliberations: no one would like to see that everyone has precommitted himself to defection. Furthermore, there is no way to determine whether it is rational to adopt a strategy of absolute precommitment, since the equally strong agents will all go through the same deliberations whether it is rational to adopt this or that strategy. In the case of **CG** the indeterminacy due to multiple equilibria creates a problem of instability. That is, when deliberating what option to choose, supposing that the others will cooperate, it pays to defect, but supposing that they will defect, one should cooperate. In the case of equally strong agents, an agent cannot get a clear vision of what to do.

Indeterminacy, no doubt, creates a serious challenge for the explanation of free riding, but the problem with the **CG** is that the eligible solutions do not strictly speaking represent the problem of

---

[27] Very powerful arguments in favor of the **CG** are provided by Taylor and Ward (1982), Taylor (1987) and Tuomela (1988, 1992).
[28] This is true also of the **BS.**

unsatisfactoriness, which is assumed to be a crucial feature of free riding. That is, the eligible solutions of the **CG** are not *jointly* dominated by the collectively acceptable strategies. Another problem is that since the **CG** does not represent a problem of coordination, there is no *collective* reason for cooperation to emerge. Consequently, there would be no option for free riding, either.

Metaphysically, the distinction between the types of agents would mean that the society is a project of risk-averse agents that are abused by a class of daring persons that do not have to involve themselves in the common project. However, it does not seem very plausible that free riders typically force the others to contribute. I mean that it is not plausible to think that the risk-averse agents would join forces in a way that the desires of the free riders would be satisfied and all would be happy with this arrangement. When the division between the types of agents is ruled out, the best option that an agent has in the **CG** is to randomize his action. That is, randomizing maximizes the expected utility, and it yields a stable solution to the problem arising from indeterminacy as well. However, randomizing means bad news in accounting for free riding. The emphasis is no longer on free riding, or on defecting, but on clear and direct maximization of expected utility, which may or may not come in the form of defection. Furthermore, randomizing means giving up the project of collective rationality in the sense of trying to find a coordinated choice. Thus, the **CG** is inadequate for representing the free rider problem, as all singular models appear to be.

It seems that the main problem in accounting for free riding is to provide a model that contains the elements of conflict and coordination in coexistence. On the one hand, free riding arises from the conflict between individual and collective rationality in a sense that it yields unsatisfactory outcomes. On the other hand, an incentive for free riding cannot arise, or at least it is not rationally conceivable, if the credible expectations of the provision of the collective good are missing. As a result, the expectations of the coordination of actions must be eligible if free riding is to be considered as a rational option from an individual's point of view.

2.2.4 Free Riding and Collective Action

The above discussion proposes that collective action has to do with coordination of actions and suitable expectations that the agents involved will take part in the efforts of reaching the collective goal. In short, collective action requires interdependence of expectations and corresponding actions. Free riding breaks the harmony of fully coordinated collective action. On the one hand, it is motivated by individual rationality that conflicts with collective rationality, but on the other hand free riding requires successful collective action. Furthermore, free riding represents a threat to collective action, and in part because of

this threat each agent that is involved is expected to take part and not to free ride. Accordingly, explaining rational collective action so that free riding fits to the picture requires an account in which conflicting expectations may coexist without this being a case of irrationality, e.g. self-deception. This means that the explication of free riding faces serious challenges that cannot be settled by referring to singular game theoretical models. This feature of the problem becomes clear from the following analysis provided by Raimo Tuomela (1992).

Tuomela's analysis is fruitful from the point of view that it basically operates with the notions of expectations and beliefs that rationalize the move of free riding.

(FR) A member A of a collective G *intends to free ride* relative to a public good produced by a joint action X if and only if

(1) A intends to defect (viz. not to contribute or do his part of X).
(2) A has a belief to the effect that the joint action opportunities for the performance of X will obtain, especially that at least K members (or a sufficient number of members required for the provision of the public good produced by the performance of X) contribute (or do their parts).
(3) A believes that he ought to participate in the production of X and that there is (or will be) a mutual belief among the full-fledged and adequately informed members of G to the effect that the joint action opportunities for the performance of X will obtain, and to the effect that each full-fledged and adequately informed member ought to contribute.
(4) A believes that he will gain more by defection than from contribution if at least K agents contribute, where K is the minimal numbers of agents capable of jointly performing X.
(5) A believes that the outcome resulting from all the agents contributing is better than the outcome when all defect.
(6) A believes that his defection involves a cost (possibly nil) to the contributing members of G. (1992, p. 174)

Condition (1) can be said to rule out the possibility that free riding would take place under condition of the coordination equilibrium, since defection does not fit the picture with successful coordination and reciprocal expectations that this coordination would take place at coordination equilibrium. On the other hand, conditions (2) and (3) point to the elements of coordination and suitable expectations of each other's taking part to the collective project. That is, an agent believes that the situation is such that when actions are suitably coordinated, a collective good can be obtained, and that the agent himself is expected to take part in the plan. Condition (2) refers to the rational conceivability of the cooperative strategy of providing the good by stating the condition under which providing it is rational. Especially, it can be said that the problems of coordination are resolved in the sense that the provision of the collective good can be expected under the given circumstances. Condition (3) refers to the reciprocal expectations under circumstances in which the agents are trying to coordinate their actions, and especially to the belief that the agent *A* himself is expected to take part as well. Condition (5) refers to an incentive for joining forces in order to provide the collective good.

However, the conditions do not give any promise of providing the desired outcome by referring only to the present game structure. This conflicting feature is supported by condition (4). Condition (4),

in fact, states that an agent maximizes his expected utility by defecting. It says explicitly that *A* will gain more by defecting than by contributing, and it says implicitly that it pays to contribute *if and only if* K – 1 agents have contributed or are going to contribute. Since it is virtually impossible to realize whether one currently faces such a situation, it pays to defect. So, condition (4) rationalizes condition (1). Consequently, an agent must believe that he does not belong to the group that is crucial for providing the collective good.

Since each agent may see himself from *A*'s point of view it is clear that free riding may endanger the provision of the collective good. Tuomela's analysis makes the element of conflict clear. It also makes the element of coordination clear. However, the analysis does not succeed in explicating how a model that contains the given elements should be constructed. It does not show what kind of mechanism justifies the suitable beliefs and expectations toward collective action in coexistence with intentions to free ride. Different game theoretical models give different answers to the questions about the motivations for free riding, but not one of these models succeeds correctly in pointing out the general mechanism that a free rider follows in his deliberations. When the agents are considered rational in the sense of maximizing one's expected utility, the choices an agent makes describe rationality rather than motives of free riding in most situations. Sometimes the agents defect and sometimes they do not. The individual models do not tell whether this or that move of defection represents free riding even in a case of providing a collective good. The only thing that says something about the motivations for free riding is the willingness to benefit from other people's efforts in a way that is not expected by those who contribute. Furthermore, the only way to explain free riding is to construct a model in which those who contribute may coexist with free riders, although the beliefs and expectations of the corresponding agents seem superficially conflicting. This is a task that requires an elaborate study of the sources of the problems of collective action as well as a reassessment of rationality and nature of rational agents. Ultimately, this means a reorientation of game theory by introducing a novel class of games, viz. funk games.

2.3 Reassessment of the Problems of Collective Action

The standard view of the problems of rational collective action puts the emphasis on direct maximization, holding that the problem arises from the clash between individual and collective rationality, and it is the individual rationality that predominates in the pursuit of resolution. Dominance of a certain option and unsatisfactoriness of the outcome are the main obstacles in the way of realizing

collective rationality. Collective rationality is, according to this approach, reduced to whatever a collection of arbitrary individual interests may provide.

However, as the study of free riding already suggests, an account of the problems of rational collective action requires a more sensitive approach to collective rationality. Especially, it is required that coordination of actions be rationally conceivable. It is also clear that the problem of unsatisfactoriness is not the only source of the problems of collective action. In addition, it should be evident that the direct way is not always the best. The *direct* maximization may itself be a source of problems from the point of view of implementing collective rationality. Instead, coordination of actions emphasizes the idea of *indirect* maximization. The idea is simple. In order for it to be meaningful to discuss problems of collective action, rational collective action itself must be somehow reasonably conceivable. When it is, collective rationality is implemented by successful coordination on the part of the agents. The maximization of expected utility may not, then, take place directly, but after suitable ways of implementing the concordant moves are found. This means that the agents are able to deal with the problem of indeterminacy that are difficult for the standard game theoretical accounts to handle.[29] Following Hampton (1987), then, the question of rational collective action may be set in two alternative ways: "How do we cooperate?" and "Do we cooperate?" From the point of view of collective rationality the first question, and consequently the problem of indeterminacy, seems primary, although the second question is the one that introduces the problems that may be terminal for collective action. The main problem of the present approach is, then, how to fit the distinct sources into the same picture.

My answer to the main problem is the introduction of funk games. I shall claim that they provide the means to fit the problem of unsatisfactoriness and indeterminacy into the same picture. This will be done by introducing the possibility that social habits can prosper in circumstances of conflict. Funk games react to the first kind of questions presented above and are able to resolve problems of indeterminacy. However, the games that react to the second type of questions, viz. games in which direct maximization by means of deliberation of a choice predominates, may nest in the funk games. This is the way, I think, the problems of collective action that are associated with the problem of unsatisfactoriness, especially the problem of free riding, can be properly understood.

The emphasis on collective rationality, however, requires a reorientation of the approach to rational behavior and collective action. One thing that is required at this point is to change the focus from direct maximization to indirect maximization. That is, the emphasis must be put on coordination, or reciprocation in expectations and actions. Another issue, to be considered in the next chapter, is the

[29] Cf. especially Harsanyi & Selten (1988).

reassessment of rationality and rational agents. Furthermore, rather than dealing with agents that before each act deliberate over alternative choices, the emphasis is on agents conforming to habit. The idea of moving along a path may clarify the difference between choice and conformity. First and foremost, however, the emphasis of collective rationality stresses the corrective aspect over the troubles that individual rationality runs into. The starting point is the problem of indeterminacy and how rational agents may deal with it.

2.3.1 The Other Way Round: It Takes Two to Tango

Coordination of actions requires making interdependent moves. That is, in order to act collectively in the sense of coordination, the agents must take into account each other's expectations and actions in such a way that concordant moves may be made. Often, especially when the agents face problems in coordinating their actions, the problem of indeterminacy is involved. If the problem of unsatisfactoriness emphasizes individual rationality in composing collective action, the problem of indeterminacy leads to collective rationality. Since individual rationality does not seem to manage with the problems of indeterminacy, the resolution must be sought from elsewhere. Collective rationality comes into rescue. What cannot be reached by separate individual deliberations is perhaps within reach by focusing on concordant moves.

A great deal of collective action is not so much concerned with conflict between individual and collective rationality, but instead with formation of suitable expectations and making recognizable moves according to these expectations. In the corresponding situations signaling of intentions and mutual dependence in reaching the desired outcome play the main roles.[30] The game of pure coordination, **CC**, provides a textbook example (cf. Figure 2.3.1) of the problem at hand.

	A	**B**
A	(1, 1)	(0, 0)
B	(0, 0)	(1, 1)

Figure 2.3.1 the game of pure coordination

[30] Cf. Schelling (1960), Chapter 4, for the first analysis of this topic.

The normal form presentation of **CC** (as in Fig. 2.3.1) clearly shows the source of the problems that rational agents face. There are two equally preferred coordination outcomes, but no *endogenous* means to figure out which one to choose. If the information provided by the game matrix is all that the agents have, and supposing that no pre-play communication is possible, the players will have no means to coordinate their actions. That is, no reciprocation in actions and expectations is possible solely from an individual's point of view. It requires at least two agents to try to calculate what the other is doing and to have common knowledge of this. The rational agents face the problem of indeterminacy.

However, under the given description of the game the players have no other option than to randomize their strategies and hope that this leads to a satisfactory outcome. In a sense, the players try to guess the others' choice, and then make a corresponding choice. Paradoxically, however, aiming at a satisfactory outcome in this way leads to the problem of unsatisfactoriness. However, if the agents were able to coordinate their choices they would do better by randomizing the choice than by direct maximization. This seems like the end of the story and collective rationality remains unaccounted for, unless the elements of *recurrence* and *continuance* are brought in. They seem like necessary conditions for collective rationality. This is, in fact, a natural amendment to the approach if the implementation of collective rationality is seen as a matter of the emergence of social habits or customs, as is the case in this essay.

Now, when agents recurrently face the problem of coordination, as illustrated in the **CC**, reciprocation of actions and inducement of expectations become available. Consequently, the agents may raise hopes for coordinating their actions in order to secure the maximization of their expected utilities. The question 'how do we cooperate?' becomes current. Conforming to collective rationality may have better prospects than deliberating over the alternative options.

Furthermore, the iteration of the **PD**-type problems seems to offer hope for reaching a collectively rational resolution in the sense of coordinating actions under circumstances in which conflict predominates as well. Then, however, reassessment of rationality and rational agents require a more detailed discussion. This is due to the fact that ideally rational agents are not in the position of having to give up a reliable mechanism of extending their scope of beliefs. Consequently, they would induce backwards the *admissible* and *unique* resolution to the problem of deliberating the best choice under the given circumstances, and the emergence of collective rationality as well as the problem of free riding would still be unaccounted for. But then we would have to tolerate the counter-intuitive resolutions that postulating ideal rationality leads to. It is worth noting that even if the desperate outcome of a single-shot **PD** could be tolerated just because the structure of the situation does not give any other opportunity, under the circumstances of recurrence the resolution to which admissibility and

uniqueness lead into is harder to accept. This observation holds especially when ordinary agents are involved. This is due to the apparent opportunity of coordinating actions even for a little while. But, as has already been made clear, the reverse of the resolution to the problem of this counterintuitivity is the problem of indeterminacy.

Despite the threat of indeterminacy, under dynamic circumstances in which the ordinary agents expect to face a net of recurrent interactions, unsatisfactoriness turns out to be a problem that requires immediate resolution because of the counter-intuitivity that prolonged unsatisfactoriness leads to. Standard individual rationality seems to be inconsistent with plausible human behavior in recurring situations. Then, according to the account that takes plausible human behavior for granted when describing rational behavior, randomizing does not count as an eligible solution when a certain confidence and predictability of actions and expectations would provide more promising results. Then, ordinary agents may find it rational to consider other options than that leading to an unsatisfactory outcome, which in the long run becomes even more counter-intuitive. Ultimately, this means a remarkable weakening of the condition of uniqueness.

However, when the demand of uniqueness is considered to be incoherent, the problem of indeterminacy is brought in, and ideal individual rationality turns out to be incapable of providing a rational resolution in any form, unsatisfactory or otherwise. Then the answer must be sought from elsewhere, for example from collective rationality, which basically takes place by inducing reciprocal expectations based on which the agents find the right course of action. The problem of collective action is, then, to construct a network of reciprocal expectations, beliefs and concordant actions.

2.4 Concluding Remarks

It seems evident that the standard game theoretical accounts that hold that collective action and its problems are basically questions of direct maximization are mistaken.[31] For instance, the standard accounts do not succeed in explicating free riding. The nature of rational collective action cannot be properly understood from the point of view of emphasizing individual rationality alone, and holding that the problems of collective action are basically questions of unsatisfactoriness does not succeed in accounting for rational collective action at all. That is, since under given description there is no collective action to discuss, no corresponding problems can emerge, either. Consequently, e.g. free

[31] Recent studies, e.g. Hampton (1987), and personal communications with R. Tuomela (the author of *A Theory of Social Action* (Reidel (1984) and *The Importance of Us* (Stanford UP (1995)) have convinced me that collective rationality

riding is left unaccounted for. In addition, reducing the problems of collective action to the question of unsatisfactoriness evades the main issue, namely the question of rational collective action.

Besides evading the issue of rational collective action the reductive account appears to be unsatisfactory even for accounting for individual rationality, since it turns its back on the possibility of the corrections that the collective rationality may provide to it. Especially, the reductive account is committed to recommendations that are not acceptable on intuitive grounds.[32] This unacceptability is due to the inadequacy of the standard account. Although under the ideal circumstances the standard account succeeds in providing a reliable mechanism for extending the belief system in a coherent and consistent manner by means of admissibility and uniqueness, in the real world the agents' capacity of making the necessary recursions may be, and in fact is, limited. Then, what holds for problems presentable in a normal form does not always hold for rational behavior in social interaction. Yet, rational behavior is something that we may face in our everyday lives. Even if we were not capable of taking every contingency into account when taking a course of action, the move may be thought of as rational. Consequently, it is evident that a reassessment of rationality is required. I shall investigate this issue in the next two chapters. Thereafter, I shall turn to the issue of collective rationality that is brought about by means of e.g. social habits. Social habits, however, cannot be naturally explicated in terms of rational deliberation over alternative choices alone, that is, in terms of standard game theoretical notions. Rather, social habits originate through conformation to patterns. An interesting feature of conforming to a pattern is that the pattern itself in part determines what the rational course of action is. Focusing on the maximization of the expected utility alone is not enough. From this perspective, collective rationality adds to individual rationality, and it is evident that collective rationality is not reducible without a residual to individual rationality. This is a basic hypothesis for reconstructing a model of collective rationality, and in part for this reason I find the introduction of funk games as games of collective rationality fascinating and helpful in explicating the problems of collective action. The central argument for this effect is given in Chapter 5. As this is a task that requires the reassessment of rationality, I shall now turn to the issue of folk rationality, viz. rationality of ordinary agents, in contrast to ideal rationality of superior beings.

---

cannot be explicated in terms of individual rationality alone. Nevertheless, I believe that without the notion of individual rationality collective rationality cannot be explicated at all.

3 THE FOLK CORE OF RATIONALITY

3.1 The Foundations of the Account

The folk view of rationality holds that rational action is determined in terms of suitably organized beliefs and desires. In order to check whether the requirements for rationality are fulfilled, the beliefs and desires are usually measured by means of probabilities and utilities. Especially, the epistemic interpretation of probability developed by John Maynard Keynes[33] and improved by Frank Plumpton Ramsey[34] is of interest. This approach holds that it is natural to consider probabilities as degrees of beliefs when rational action and evaluation are at stake. In short, when discussing rationality we are dealing with the *logic of partial belief*,[35] and a natural measure of the partial belief is probability as a degree of belief. Furthermore, in order to construct a general and exact theory of the degrees of beliefs, Ramsey adopted a theory *of satisfying desires at a maximum level* – the theory of utility. He considered this theory to be a good approximation of truth in explaining human behavior in folk terms. When desires are ranked by means of the theory of utility it is possible to introduce degrees of certainty in terms of expectations, since when it is assumed that the agents aim at satisfying their desires at the greatest level, a certain behavior can be expected from these agents.

Furthermore, Ramsey emphasized that rational agents must follow certain recursive rules in making inferences, and that there are limits to the recursions of beliefs, since there are limits to the capacities of observations, memories and making inferences. Ramsey held (1931, p. 184) that the requirements of perfect inferential skills, recall and knowledge are "too high a standard to expect of mortal men". So, "we must agree that some degree of doubt or even of error may be humanly speaking justified". According to Ramsey, then, “the perfect certainty and the truth of an opinion are ideal more suited to God than to man.[36] These lines, I think, express clearly enough the idea of boundaries in the capacities of ordinary rational agents. They suggest that there are constraints that every model of epistemic rationality must take into account. Especially, ordinary rational agents are supposed to be finite and fallible. It will become clear that for this reason, the best we, the ordinary agents, can do is to conform to useful habits - reliable mechanisms - that guide us in satisfying our desires. That is, not

[32] Cf. especially the chain-store paradox introduced by Selten (1978), and its offspring, e.g. the backward-induction paradox.
[33] Cf. Keynes’ *Treatise on Probability*.
[34] Cf. e.g. Ramsey’s ‘Truth and Probability’, first published in 1926. Ramsey (1931).
[35] Cf. Ramsey (1931), p. 166.

beliefs alone, but combinations of beliefs and desires supplemented with certain general habits and rules of inference explain (rational) human action.

The folk psychological trait in Ramsey's work, belief-desire psychology complemented by characteristically rule-governed ways of making inferences, makes Ramsey's approach functionalistic in the sense proposed e.g. by Lewis (1983). The whole of beliefs, desires and habits are what compose rational action. That is, we are dealing with a holistic account in which rationality, in a sense, supervenes on beliefs, desires and habits. The functionalistic tone of the approach is powerfully expressed by presenting the idea of beliefs as maps.[37]

The works of Ramsey, especially 'Truth and Probability' written in 1926, and 'General Propositions and Causality' written in 1929, are instructive for understanding the current discussion that has certain epistemic and belief logical tones and that operates with incomplete agents. In the first of these essays Ramsey explicated the idea of probabilities as degrees of beliefs and the idea of presenting desires as utilities. Here, then, Ramsey laid down the cornerstone of game theory. In the latter of the writings he connected the above ideas with a common sense view of the mental that, in a sense, anticipated the modern functionalist views.[38] I believe that this functionalism in Ramsey can and should be applied in the present account of folk rationality.

3.1.1 Rules and Habits

Choices between gambles, or as Von Neumann and Morgenstern (1944) called them, choices between lotteries, connect beliefs and desires suitably for generating recommendations for rational action. These lotteries or gambles yield measures to both beliefs and desires. The main point of yielding the measure is the idea of introducing a certain habit of inference that guarantees that the induced beliefs and recommendations of action are coherent and consistent. Such a habit of inference provides a way to reach a conclusion that the agents can be relatively certain of. That is, in game theory it is required that beliefs are, among other things, coherent and consistent. Or, to be more precise, the whole of beliefs and desires, that is the belief-desire system, is coherent and consistent.

Consistency as a requirement of the formal logic does not depend on the meaning or contents of beliefs. It could be claimed that the purpose of logic, as Ramsey (1931, p. 191) put it, is simply to ensure that our beliefs are not self-contradictory. This is, however, not enough since we do not want

[36] Ramsey (1931), p. 194.
[37] A number of authors, for instance Armstrong (1973), Nelson (1982), Lewis (1983) and Sorensen (1988) embrace the idea of beliefs as maps as an apt allegory of the mental processes.

our beliefs to be consistent only with one another, but also with facts. This means especially that rationality is inextricably inductive. According to Ramsey, "it may humanly speaking be right to entertain a certain degree of belief in truth of our propositions on inductive or other grounds". The idea becomes clear when the following passage of Ramsey (p. 192) is observed: "consistency combined with observation and memory is frequently credited with the power of leading to truth". That is, we require that our beliefs are coherent, as well. Consistency and coherence are requirements that dispose us to make "correct" inferences.

Thus, the present account suggests a system that is not reducible to formal logic, since this system appears to be tied to the capacity of human beings to observe and remember, and inferences from them to further beliefs may be and often are inductive. Ramsey aimed at explicating what he called a "human logic"[39]: a rule or habit that people believe to be reliable in the process of making inferences. Hence, according to Ramsey, the human mind works according to general rules or habits, and among the habits of human mind induction has great importance. Adopting inductive means is reasonable. That is, induction is a useful habit on the path of satisfying our desires in an effective way. For instance, the maximin-principle: "in unprofitable situations, choose an alternative that minimizes expected losses", or the best-reply principle: "always choose an alternative that is the best reply to the opponent's expected choice" are examples of rules and habits that aim at guaranteeing consistency and coherence of the inductions. Other corresponding principles are those of uniqueness and admissibility, which are studied in more detail later in this chapter.[40]

Ramsey called the habits of inference *general beliefs*. General beliefs, whatever rules of inference they may have, belong to the requirements of rationality. Appreciating the requirements of rationality can be said to guarantee, if not the truth of conclusions, then at least satisfaction of one's desires. So, these requirements have a function in an agent’s rational behavior just as beliefs and desires have their functions.

However, given the above characterization of rationality, an example in which the contents of an agent’s beliefs and desires may be considered as eccentric is perfectly conceivable and, still, the agent in question may be considered to be perfectly rational, as long the rules of inference are observed. Consistency does not rule out eccentricity. Only in the case in which the beliefs are known to be false they should be eliminated in the name of consistency, provided, of course, that it is irrational to hold that 'I believe that *p* and I know that *not-p*'. It is also clear that the range of application of the rules and

[38] Mellor (1990), p. xviii. Cf. also Block (1980).
[39] Cf. also Nelson (1982).
[40] Cf. section 3.2.3.

habits is not unlimited. For instance, the dominance principle, although it is generally considered as uncontroversial, is very restricted in its applicability. Also the limits to the depth of beliefs of ordinary agents set boundaries to applying the rules and habits in making recursions of the future options.

Consequently, a further requirement is needed: the approximate correspondence to reality. This requirement, call it a *general desire*, that we succeed in satisfying our desires when acting according to our beliefs warrants us to eliminate inferences that do not lead to satisfactory results, despite the fact that they are in accordance with consistency or coherence. Of course, inconsistent beliefs never qualify. The general desire, thus, disposes us to require that the habits of inference we possess do function in a way that the satisfaction of our desires according to our beliefs is expectable in some law-like manner, deterministic or stochastic. In other words, general desires provide a loop back to the present beliefs and desires, and insight as to how the latter are connected to the general beliefs. The basic idea is that whenever one makes an inference, one does it according to some rule or habit, which basically fulfils the requirement of consistency and coherence. In addition, more importantly, the inference one makes is expected to correspond to reality, or at least serve to satisfy one's desires. Furthermore, the idea of general beliefs as habits or rules anticipates the idea that the inferences the agents make are recursive and that they follow certain relatively simple rules. The same holds not only for inferences, but also for observations and memory as well. However, their inductive nature implies that the absolute certainty is ultimately impossible to attain, as Ramsey noted.

3.1.2 Belief Maps

The apparently demanding requirements of consistency and coherence become understandable when Ramsey's idea of 'beliefs as maps' in 'General Propositions and Causality' is studied. [41] These requirements must be considered as certain minimal conditions. That is, our belief systems may not be incoherent and inconsistent, even though they may be inaccurate in certain respects. It is natural to think that the requirements of consistency and coherence belong to the rules that enable us to extend, so to speak, the original belief map.[42]

Seeing beliefs as maps must be considered as a metaphorical and vague way of stating the issue, but it illuminates some basic ideas. The maps we are dealing with are incomplete and finite. They may contain errors, blank spaces and even fantasies.[43] They may even involve contradictory representations

[41] First published in The Foundations of Mathematics and Other Logical Essays (1931).
[42] Of more detailed discussion of this issue, cf. e.g. Nelson (1982), Ch. II. Nelson speaks of rules of mind.
[43] Cf. blindspots and inexact knowledge.

of the world.[44] This idea, brought back to light by Armstrong (1973), is not of interest merely from the game theoretical point of view but it is of interest for the functionalist account of the mental, as well. 'Beliefs as maps' means essentially that beliefs are to be seen as action guides. The habits of inference we entertain dispose us to make corrections to the picture given by the original belief map. So, as "beliefs are maps of the world in the light of which we are prepared to act",[45] it is reasonable to require that they are, at least, *not in*consistent and *not in*coherent. For when inferring what to do, we want to expect that we can rest assured that our beliefs will guide us approximately to the right track in satisfying our desires. The map points out a path that is good to take.

These are the ingredients of which a folk notion of rationality can be composed. Let us now consider the approach of folk rationality in more detail.

3.2 The Portrait of an Ordinary Rational Agent: the Background Assumptions

The present account takes into reconsideration the classical game theoretical view of rational behavior, especially of what can be called the eductive approach. [46] Thus, I focus on the interaction between rational deliberators rather than on evolutionary mechanisms, although while recognizing this distinction, the current account tries to bring these two approaches closer to each other. My treatment of the eductive approach calls for several specifications to the classical analysis, although the foundations of game theory are embraced. I also aim at bringing the approach close to its folk core, by which I understand an approach that corresponds to the analysis set forth by Ramsey.

Furthermore, it is widely held that the current game theoretical analysis is in its core a Bayesian project. This is especially true of eductive game theory based on subjective probabilities. Bayes' formula shows how *a priori* beliefs are transformed into *a posteriori* ones in the rational deliberation process. For instance, the works of Aumann (1974, 1987) and Skyrms (1990) emphasize the Bayesian nature of the theory. On the other hand, Von Neumann and Morgenstern remained silent on this evident trait in *The Theory of Games and Economic Behavior* (1944) in the dawn of the game theoretical discussion. However, acknowledging the Bayesian core of rationality does not need to commit us to Bayesianism in general. In fact, Bayesianism as a general theory does not give an accurate account of rationality.[47]

Although I am proposing a non-Bayesian account of rationality, Bayesianism cannot be totally

[44] Cf. e.g. the instant recommendations of collective rationality and the result of the present deliberation.
[45] Armstrong (1975), p. 4.
[46] The term 'eductive', to my knowledge introduced by Ken Binmore, derives from the concept of 'eduction', singular induction, a term suggested by W.E. Johnson. The notion naturally emphasizes the inductive character of rational deliberation inherent in game theory.

erased from the core of this account. Its Bayesian core can be expressed e.g. as in the standard game theoretical accounts. That is, a rational agent should choose an option that *maximizes the expected utility*. However, since we are living in an imperfect world and we are equipped with incomplete capacities, Bayesianism as such does not adequately describe rationality. For instance, because of the shortage of capacity and limits in depth of the beliefs, we put great effort into adopting habits and learning routines that facilitate our everyday living. Furthermore, we are doing so in part because this is the way in which our desires are satisfied most effectively. So, in the present study of rationality the frame is non-Bayesian, even though the core is tied to a central Bayesian idea of maximizing expected utility.

3.2.1 Boundedly Rational Agents with Incomplete Knowledge

These observations of the philosophical background of the present account state quite explicitly that rational action is tied to boundaries in making inferences, and consequently to the fallibility of agents, and, furthermore, to a certain robustness of these rational agents. It is evident that if we are to treat rationality as a property of finite agents with restricted capacities, the standard postulates of complete knowledge and perfect rationality must be given up. Since it is plausible to assume that rational agents are aware of their restrictions, they are also capable of tolerating mistakes - that is, an off-equilibrium act does not always lead to condemnation of the performer of such an act as irrational. As the performers are assumed to be ordinary incomplete and finite agents, at a certain point the deliberation must stop - and at a certain point a decision must be made and ultimately put into action with no hesitation. Consequently, the depth of the capacity of recursions of beliefs is limited. As there are, without doubt, limits to the recursions of beliefs of ordinary agents, it may be ultimately impossible to verify that the choice of a rational agent is, in actual fact, the uniquely best response to what she expects the other agents to do in an arbitrary game situation, let alone the fact that the beliefs themselves may be incomplete. Binmore (1987) speaks of a "stopping-rule" and a "guessing-algorithm" that are inherent in boundedly rational reasoning processes, implying that at certain point, when the stopping-rule so dictates, boundedly rational rule followers jump to a conclusion according to the corresponding rule. [48] In other words, we are dealing with finite recursive computing systems liable to incompleteness and

[47] Cf. e.g. Binmore (1993) for a detailed discussion.
[48] Cf. also Skyrms' (1990) adaptive rule of the Bayesian deliberators.

underdetermination and, hence, liable to fallibility.[49]

It is instructive to pay some attention to the requirements for rationality of the classical approach in order to get a clearer picture of the issue. The first requirement states that everything that there is to be known about the circumstances of the agents is known. Secondly, the agents are required to be able to *completely* and *transitively* order their options according to their beliefs and desires. Ultimately, this means that the agents are equipped with ideal inferential skills.

Usually, in game theoretical analysis perfect and complete knowledge (or, information, to be more precise) of the game are separated from each other. Perfect information means that in each choice situation the players know how they are situated. Complete information means that the players know what options they and the opposite parties have in the game as a whole. That is, the players know the structure of the game.

In the classic analyses it has been a standard assumption that the requirement of complete information is fulfilled, whereas information about the current situation is allowed to be imperfect. I take the postulate of complete information to serve mainly the goal of simplicity. Generally speaking, the postulate of complete information appears to be redundant in defining rationality, and it is, as a matter of fact, highly misleading, given the philosophical foundations of the current approach presented earlier in this essay, since it does not plead for rationality as such. Game theorists have also given attention to the role of complete information in their work of the models of rational behavior.[50] The incompleteness of the capacities of our observations and memories, as well as the way in which we infer from prior beliefs to further beliefs are made, suggest quite another picture of rationality than that suggested by the models that operate with the perfect and complete knowledge.

However, when analyzing certain situations, it has proven useful to make certain idealizations, such as postulating perfect and complete knowledge. It should nevertheless be kept in mind that the idealizations in accounting for the logic of mind or rationality serve only some preliminary purposes, in this case especially the purpose of facilitating the revelation of the general mechanisms of rational deliberation.

Although the postulate of complete knowledge about the structure of the game is redundant for the purpose of defining rationality, some information of the structure and about the other player(s) is necessary for making rational choices, since by rationality we mean, roughly put, the product of beliefs, desires and action. Moreover, the beliefs the agents entertain are expected to cohere to the facts of the

[49] Nelson (1982) discusses computational rules of belief (Chapter II) at length. E.g. Anderlini (1990), following the ideas of Binmore (1987), brings recursive functions to the game theoretical discussion.
[50] Cf. e.g. Harsanyi & Selten (1988) for a discussion of incomplete knowledge.

situation the agents are facing. If they do not, the satisfaction of desires does not take place by virtue of referring to the correct reasons.

For instance, in order to *coordinate* their actions the players must have common knowledge not only with respect to rationality, but also to the relevant part of the game, namely to the part in which coordination of action is possible. So, rationality is connected to knowledge or beliefs, but it is not required that the knowledge or beliefs be complete, although the more complete the belief system or knowledge of agents, more likely it is that their desires are satisfied by virtue of referring to the correct reason. A notion of local completeness - a range of mutual beliefs - may turn out to be a useful specification of the postulate of completeness when discussing e.g. coordination. By this I mean that the players may have private information concerning the game, but when they want to play safe, that is, make the choices they are expected to make, they make their choices on the grounds of the information that is commonly known by the players.[51] The information about the commonly known outcomes of the strategies is locally complete. Thus, the private information may be considered as virtually irrelevant for the present choice situation, even though making a choice based on the private information is still an option for the players, and it might even be rational from the individual point of view. Recall that rationality is in part composed of certain inductively "valid" inferences - habits that guarantee satisfactory results on subjective grounds connected with raw beliefs and desires.

The assumption of perfect or ideal skills and capacity of making inferences, in short, the assumption of perfect rationality, is another standard assumption of the classical approach. On the grounds that we are dealing with ordinary incomplete and finite agents, the assumption of perfect rationality is even more problematic than the assumption of complete information. After all, there are situations in which agents can be said to have complete knowledge, whereas the notion of perfect rationality is usually applied in a manner that makes it something that is more suitable for superior beings - a regulative rule that no ordinary intelligent agent can *de facto* satisfactorily follow. The notion of perfect rationality means not only that an ideal agent has a belief system that entertains e.g. coherence and consistency etc., but in addition that the ideal agent is supposed to be equipped with a perfect computational ability that the agent is actively using, as well. As rationality is understood as a property of the whole of beliefs, desires and actions, certain constraints to the computations of agents are required. If there are no such constraints,[52] the possibility of infinite regress of computations of

[51] I use the term ‘common knowledge’ in a somewhat loose sense. I do not want to discuss the differences between common belief and common knowledge here.

[52] E.g. a stopping-rule that consists of a simple mechanism of marginal product of making the extra inferences that go beyond the point of ordinary capacities - that is beyond the limit of the depth of beliefs that agents entertain in their everyday living.

"ideally rational" agents becomes a serious threat to the pragmatic appeal of the model.[53]

In contrast to the picture given by the model that operates with the assumption of perfect rationality, we people are faced with boundaries - there are limits in the depth of our beliefs and in our capacities of processing those beliefs. We are finite "computers" capable of only a limited number of recursions of our beliefs.[54] Yet, we are inclined to consider ourselves as rational beings - beings whose behavior is conducted by beliefs, desires and intentions that are ideally organized in some effective way. Thus, bounded rationality seems to be a central implication of the philosophical background theory of the current account. The inferences the agents make are inextricably inductive, the capacities of the agents are limited, and the recursions of beliefs must stop sooner or later in order for rational action to take place.[55]

3.2.2 Fallibility

As we are dealing with incomplete finite agents, it sounds natural to expect them to be liable to err, as well. Mistakes on part of the agents are not necessarily products of irrationality, although a token act or sequence of acts may seem irrational in a case in which it is based, in fact, on an error due to incomplete skills, capacities, knowledge, etc.[56] Besides restricted skills and capacities - that is, boundaries - trembles, blindspots and inexactness of knowledge may lead to apparently irrational behavior according to the classical approach. In the classical approach the agents are assumed to be perfectly skillful - rational agents do not tremble, nor do they make any inferential mistakes, and the solutions of the games are restricted to profiles that under certain circumstances have been found by some authors to be unintuitive - thus ineligible. One way of introducing the possibility of intuitively more acceptable results and deviations from an equilibrium path is to allow that mistakes can happen. That is, ordinary rational agents are liable to err, e.g. engage in sequences of strictly speaking irrational acts.[57]

To allow the incompleteness in skills, the present account allows that e.g. random errors (i.e. trembles) may happen, and according to Selten (1975) they may even have beneficial effects, at least in sense of showing the inadequacy of the standard solution concept of *Nash equilibrium*. The main point

[53] E.g. Anderlini (1990) defends a view that human decision makers cannot compute functions that are not recursive.
[54] The word "computer" must be understood in the sense in which e.g. Ramsey and others from the "pre-computer" era used it.
[55] Cf. the discussion in Rescher (1988) for a more detailed analysis of rationality in beliefs, actions and evaluation.
[56] I take it that irrationality is expressed in terms of *akrasia*, self-deception, and wishful thinking. Being mistaken does not necessarily fall into any of these pure examples of irrationality.
[57] To be more precise, acts due to errors do not as such correspond to irrationality, but to a certain non-rationality, since they do not necessarily come from self-deception, wishful thinking or *akrasia*, or from any other source of irrationality.

of allowing the possibility of trembles is to introduce the notion that although the behavior of the players may be unexpected in the sense of uncorrelated and stochastic errors, it is still possible to appreciate the players as rational. For instance, Binmore (1987) discusses an example in which a perfectly rational agent is supposed to make a sequence of irrational moves.

However, the discussion of irrational moves of rational agents that goes in terms of counterfactuals in the light of Lewis (1976) appears to be somewhat incoherent, since perfectly rational agents do not make any mistakes, nor do they make any irrational moves. Such considerations are better understood, and the problems of incoherence and inconsistency of the analysis are avoided, if the focus is set on the fact that boundedly rational ordinary agents are liable to mistakes. The first effort to develop a trembling-hand explanation was made by Selten (1975).

Another source of errors, in contrast to the stochastic ones, may be found inside the belief systems of the agents.[58] Besides making trembles, our belief systems may suffer e.g. from blindspots; there may be plenty of inferences actually never made that one could have made, and there may be a lot of features never actually taken into consideration that one should have taken into account. In short, there may be some epistemic failures (or, internal systematic errors) in one's belief system, as well as limitations due to boundaries of one's capacities. The present account suggests that in order to make the explanation that goes in terms of rational behavior more responsive to common sense, it must be dealing with boundedly rational agents, whose belief systems may contain blindspots, and whose actions may be trembled, and, furthermore, whose wishes may be blurred. These specifications to the notion of rationality make the notion inextricably fallibilistic.[59]

3.2.3 A Reliable Mechanism

Fallibility is, thus, a built-in property of the present conception of rationality, since, as the above already suggests, we are dealing with boundedly rational agents with partial knowledge and incomplete beliefs. Necessarily, then, it is humanly speaking impossible for ordinary agents to make inferences that can *a priori* be proven to be correct.[60] However, a certain approximate "correctness" (i.e. reliability) of the inferences, i.e. that a corresponding choice provides a satisfactory result, is required if the inferences are to fulfil their part in rational action, and this requirement is fulfilled by a reliable mechanism the agents

[58] Cf. Sorensen (1988) for detailed discussion.
[59] Including, among other things, boundaries in skills and capacities, and allowing the possibility of trembles and blindspots.

apply. That is, the rationality of the inferences is in part gained by correct habits of inferences, not by the infallibility of the inferrer. Unfortunately, the habits of inference cannot be shown to be correct before the damage is done, so to speak. The best the agents can do is to expect on inductive grounds that the reliable mechanism will work. Rationality is not restricted to infallible beings, such as gods, imaginary supercomputers, ultra-intelligent extra-terrestrials or whatever our imagination may produce. Because of the nature of the cited mechanism, no deductive closure generally takes place. Instead the agents follow the logic of mind that in part rests on certain rules and habits that are supposed to provide a reliable mechanism for making inferences. Especially, this mechanism is supposed to guarantee as coherent and consistent extensions of one's beliefs as is humanly speaking reasonable to expect.

In part because of their finite skills and capacities in making their inferences - deliberating whether to choose this or that option - it is reasonable to expect that rational agents apply a simple reliable mechanism that is extracted from previous deliberations and experience. Or, alternatively, it is simply passed on by the previous generations, and it is adopted or learned without true deliberation. In other words, it is a habit or convention that is conformed to and that everyone is expected to conform to in order to guarantee the coherence and consistency of beliefs and their extensions. In this way the agents in an effective way without the pains of futile reconsiderations guarantee that their inferences are sound. By conforming to such simple habits of inferences the agents are also freed to reserve their capacities for more important issues and their interaction with each other is facilitated.[61]

The basic constituents of such a reliable mechanism are provided e.g. by the requirements of *admissibility* and *uniqueness*. Admissibility and uniqueness state that no one knowingly chooses a dominated option and that the strategy the players choose is recursively dominant, respectively.[62] In the classical game theoretic approach these requirements are interpreted strictly and narrowly: the players will choose at each round of the game an option that maximizes their expected utilities in that very node of the game, and this is commonly known by the players. [63] The players expect each other to follow these requirements of uniqueness and admissibility strictly, and, thus, they expect that the game will have a determinate solution. These expectations are warranted, since the players are supposed to be perfectly rational. Thus, the requirements of uniqueness and admissibility determine an equilibrium path

---

[60] Cf. the impossibility theorem by Anderlini (1990), pp. 27-33. Note also what Binmore (1993) says about the essential incompleteness of rationality, and about stopping-rules and guessing algorithms (1987, 1988) that are supposed to make the forecasts more reliable under given circumstances.

[61] For similar ideas, cf. Bratman (1987).

[62] Cf. Bacharach (1992), p. 248. Domination of a strategy means that no other strategy can beat it, and a recursively dominant strategy means that it is at present a unique survivor of the options the agent initially had. A recursively dominant strategy is, thus, in the above sense a determinate solution to the game.

[63] Cf. the backward induction argument.

of the game, that is, a self-enforcing solution of the game. The Nash equilibrium may be considered as an example of the solution of a classical approach.[64] However, as the boundaries and incompleteness of beliefs and skills of agents introduce a certain aspect of fallibility into the account, the solutions that the agents are seeking are obviously ultimately underdetermined.[65]

The indeterminacy of solutions is not a problem that is restricted to the models of bounded rationality. It is an acknowledged problem of the classical approach as well.[66] This problem seems inevitable, since the strict and narrow requirements for solution concepts seem to lead to counterintuitive results that call for correction. Besides the problem of indeterminacy, the problem of unsatisfactoriness of solutions has intrigued philosophers and social scientists.[67] The problem of unsatisfactoriness is a clear product of the strict and narrow interpretation of the reliable mechanism. However, relaxing the standard requirements for rational inferences in favor of making intuitively more acceptable solutions eligible seems to introduce the game-theoretically speaking more serious problem of indeterminacy. Thus, either the applied model does not always guarantee a determinate solution at all, or when it does, it may happen that the solution is not intuitively appealing. A resolution to this package of problems seems to require radical corrections for the notion of rationality itself. This is especially the case when rationality is seen as a function of e.g. equilibrium. As we are dealing with agents with limited recursions of beliefs and incomplete knowledge, the need for such specifications becomes even greater.

In order to make the account intuitively more appealing, while not sacrificing the reliability of the applied mechanism, several suggestions for specifying the solution concepts have been introduced in recent game theoretical literature. For instance, Aumann (1987) holds that rational agents make their decisions on the grounds of reciprocal beliefs, thus giving up the idea of probabilistic independence of decisions. Aumann and others argue that because of the Bayesian undertones of game theory, the postulate of probabilistic independence of choices, inherent in Nash equilibria, is unwarranted. [68] Bernheim (1984) and Pearce (1984) claim that certain strategy profiles are more "reasonable" than

[64] The Nash equilibrium as a preliminary solution concept has great appeal because of its simplicity. Furthermore, Nash equilibrium can be said to be "pedagogically" virtuous - it has the same appeal that e.g. Newton's mechanics have. However, it has proven unsatisfactory in many respects - mainly because of its restricted applicability in social circumstances, but also because it ignores some operationally useful features and because it postulates some operationally redundant features. That is, Nash equilibrium is, on the one hand, too broad because it allows intuitively unreasonable solutions and, on the other hand, too narrow because of not allowing certain intuitively acceptable solutions. Thus, Nash profiles turn out to be, in a sense, non-eligible. In the light of the current account, sticking merely to Nash profiles seems to be a symptom of the stagnation of rationality. These problems arise when the Nash equilibrium is applied under circumstances that are not those of the perfect world.

[65] Cf. Nelson (1982).

[66] Cf. Harsanyi and Selten (1988) for a detailed discussion of the basic problems of game theoretical solution concepts.

[67] Cf. all the literature concerned with resolving the PD.

others, allowing the choice of imperfect equilibrium paths. Intuitively put, this means that "unreasonable" Nash equilibria can be excluded from the set of applicable strategies, while the notion of equilibrium is given a broader interpretation, as is the case with the notion of correlated equilibrium. Ponssard (1990) and Al-Najjar (1995) hold that e.g. precedent has importance in making rational decisions, introducing forward induction (**FI**). Besides discussing the refinements to the notion of equilibrium, a broader view of a solution to the games is introduced by the above suggestions. The emphasis is on the idea of self-enforceability of a path in extensive games instead of the self-enforceability of equilibrium. To the extent that the cited considerations are taken for granted, we are not tied to Nash profiles in search of solutions for decision problems and, still, we are able to identify reasonable solutions in the set of all equilibria. The cited specifications to the solutions concepts may be summarized under the label of the signal account. In order to make rationalized moves, or to choose a correlated equilibrium path, or, in general, to make forward inductions, the agents must have mutual beliefs and reciprocal expectations of each other's rationality. These, in turn, require that the agents have enough knowledge of each other's past behavior. Such properties are summarized in David Lewis' seminal work: *Convention: A Philosophical Study* (1969). The problem with the recommendations of the signal account is that they may be inconsistent with the recommendations of the reliable mechanism. The resolution of this problem can be found by introducing funk games, as will be seen later in this essay.[69]

3.2.4 Robustness

The present picture brings to surface an important feature of rationality - robustness. Being aware of the ultimately fallibilistic nature of rational inferences and accepting that rational agents are liable to err call for tolerance on the part of the agents. Especially, agents may be expected to tolerate each other's mistakes and appreciate each other's rationality despite these occasional mistakes. Tolerance up to a certain limit is a basic factor of robustness of rationality. Another characteristic of robustness is resilience. As rational agents are supposed to apply certain simple and reliable mechanisms for making inferences and they are supposed to be capable of stretching their recursions further when necessary, their rationality may also be expected to be resilient. For instance, when correcting false expectations in order to return e.g. to the equilibrium path, it is supposed that rationality is resilient. That is, the reliable mechanism helps to sustain rationality in behavior. In short, although finite and incomplete agents are

[68] E.g. Vanderschraaf (1995) and Skyrms (1990).

liable to err, they are also capable of immediately recovering from the unfavorable state of affairs that resulted from their mistaken choices when they recognize their true position. These factors, supported by authors of altering domains, point out the robust property of rationality. E.g. Michihiro Kandori (1992, pp. 71-72) requires that the rationality of a choice does not depend on the fine detail of the knowledge and he insists that the players tolerate to some extent each other's mistakes. On the other hand, Frank Jackson and Philip Pettit (1993, p. 273) see the issue of robustness as follows. "By a robust solution we mean a solution which would not be exposed as being based on the misleading appearance of a pattern in a biased sample, by the acquisition of any actual or possible new data concerning what a person does or would do in circumstances." Speaking of rational expectations of each other's behavior I take this formulation to indicate a reliable mechanism that in part justifies the solution at hand while still allowing that in actual fact the solution may be mistaken. In other words, in a case in which new information is acquired, it is still possible to appreciate as rational the choices that were made in the absence of that information. Thus, a robust solution guarantees that there always is at least one admissible (but not necessarily unique) solution to each problem of choice. Furthermore, robustness indicates a tolerance for errors and mistakes due to boundaries, blindspots, trembles etc., since it is a solution that, given the present information, can be appreciated as rational whatever were to happen, while also pointing at the resilient character of rationality because of appealing to the applied reliable mechanism. Although a solution which is based on information that in the light of the present knowledge is insufficient cannot be accused of being irrational, the reliable mechanism requires immediate correction (if necessary) to the expectations of the behavior of the players, thus indicating the resilient character of robust rationality. In short, if rationality is to be considered as something suitable for incomplete and finite beings, such as we people typically are, it should be required that rationality is a property that survives from errors in the sense that it tolerates them, while also being capable of recovering from their consequences.

Robustness, thus, brings us to the very core of a common sense notion of rationality and the theory of persons in general.[70] Not every little detail makes a difference in the resolution of problems of choice. It is the reliable mechanism by which beliefs are suitably organized. This reading of robustness introduces among other things the idea of resistance to reconsideration. We do not question our habits or conventions unless something goes recurrently and seriously wrong. We do not question our habits of inference or our raw beliefs unless something goes recurrently and seriously wrong. We do not

[69] Cf. also the discussion on free riding.
[70] This idea is supported by e.g. Lewis (1983), p. 114.

presently engage in such deliberations unless we face a novel or unrecognized and distinct situation.[71] This trait guarantees the firm execution of action under similar distinct circumstances. It may also sometimes be called stubbornness. The choice of a notion indicates the risks that are introduced by a rigid commitment to a certain habit. The risk of being driven to the wrong track belongs to the most obvious risks. The mechanical inferences may lead to a wrong track in the course of satisfying one's desires. However, once agents realize that they have gone wrong way, they should in the name of rationality start rewinding the precedent inferences. The rewinding takes place by applying the very same rules of inference as does the extension of one's belief system. Only very rarely are the rules of inference themselves taken into reconsideration, although this can happen as well. This is what the issues of tolerance and resilience are all about.[72]

4 WHAT DO YOU EXPECT WHEN YOU EXPECT RATIONALITY?

4.1 Resolving the Problems of Unsatisfactoriness

I have now gone through what I consider to be the basic elements of the folk view concerning rationality. In short, it says that in order for an action to be rational, it is among other things required that the beliefs of the agent are organized according to certain intellectually acceptable requirements, e.g. admissibility and uniqueness, that guarantee that further inducements of beliefs be coherent and consistent. When they act by such rules and habits, the agents guarantee that their desires can be satisfied in an effective way. However, the capacity of applying the corresponding rules is finite, and agents are liable to mistakes. The folk view of rationality holds that despite the possibility of errors and because of the acknowledgment of boundaries, the expectations of rationality survive apparently incoherent actions. Rationality is robust. Given the above discussion, then, what entailments does the present account have?

Before going into the full-blown account and its entailments, I shall begin with an account of the folk view reduced to its hard core expressed by the standard requirements - admissibility and uniqueness - and show what a strict and narrow version of it entails while leaving aside for a moment the issue of

[71] Similar ideas are put forward by e.g. Bratman (1987), Chapter 5. Cf. also Al-Najjar (1995) for discussion.

boundaries, etc. This is undeniably a biased way of presenting the issue. But by beginning in this way I hope to clarify the otherwise complicated issue of the folk view of rationality, the ultimate target being the reconciliation of the standard requirements while making certain proposals to make the account intuitively more attainable. Especially, the basic idea of a folk view - keeping things simple - is embraced.

The need for clarification of the notion of rationality suggested above becomes evident when studying certain options that are intuitively reasonable but which are in view of the standard requirements and the postulate of the common knowledge of rationality ineligible. In this respect, the account appears to be inconsistent.[73] In the literature this result has been considered paradoxical - an expression of the backward induction paradox. Upon closer examination, however, the paradox turns out to be a product of incoherent thinking about rationality, or so I claim. Ideally rational agents are not liable to make choices that would bring about the paradox. In contrast, boundedly rational agents are not generally in a position to make conclusions that would bring about this paradox, and even if they were, they would be able to explain the apparent paradox away by referring to boundaries, mistakes or some other corresponding cause. That is, as far as ordinary finite and fallible agents are concerned, they are not willing to give up the expectations of rationality, or the possibility of cooperation, when the standard requirements seem to provide unsatisfactory results.

4.1.1 The Backward Induction Paradox

The first task in accounting for folk rationality is to get rid of unsatisfactory outcomes, since they lead to the problem of counterintuitivity. That is, if an outcome is unsatisfactory, it seems to conflict with the aim of maximum satisfaction of desires. Furthermore, if the result is counterintuitive, it cannot represent the folk view.[74] The problem of unsatisfactoriness is generally represented in terms of the backward induction paradox.

In the recent literature the backward induction paradox is usually represented in terms of the games of the sequential Prisoner's Dilemma or the Centipede.[75] No matter which is the case, the

---

[72] Note that the habits of inference discussed above and the social habits (discussed in Chapter 5) should be kept separate.
[73] Cf. the discussions of Pettit & Sugden (1989) and Bicchieri (1989) on the backward induction paradox.
[74] It must be noted, however, that not all problems leading to unsatisfactory result represents a problem for folk rationality. Cf. e.g. a single-shot **PD**. As the simple habits of inference point unambiguously to the dominating option, there is no rationally conceivable alternative available.
[75] Besides the centipede and the sequential **PD**, cf. e.g. the chain-store paradox or the surprise-test paradox which belong to the same family of paradoxes, despite slight differences. For instance, in the sequential **PD** no cooperation may ever actually take place despite expectations, whereas e.g. in the surprise-test case the test will be held sooner or later during the given sequence.

standard requirements for rationality, viz. uniqueness and admissibility, induce a folding back mechanism that is generally supposed to provide for a reliable mechanism for making decisions.[76] Only now, the reliable mechanism leads to an unsatisfactory result. On the other hand, rejecting the reliable mechanism is incoherent, thus irrational. This is what is generally understood as the backward induction paradox. A classic presentation of the problem goes as follows:

> Your corn is ripe today; mine will be so tomorrow. 'Tis profitable for us both that I shou'd labour with you today, and you shou'd aid me tomorrow. I have no kindness for you, and know that you have as little for me. I will not, therefore, take any pains on your account; and shou'd I labour with you on my account, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security. (Hume, 1739, Book III, Part 2, Section 5)

The following game tree serves to illustrate the classic features of the centipede:

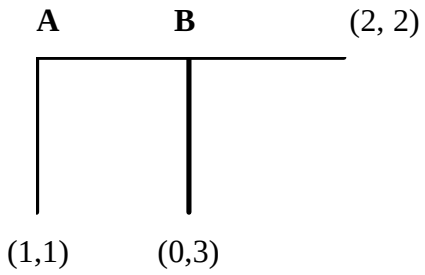


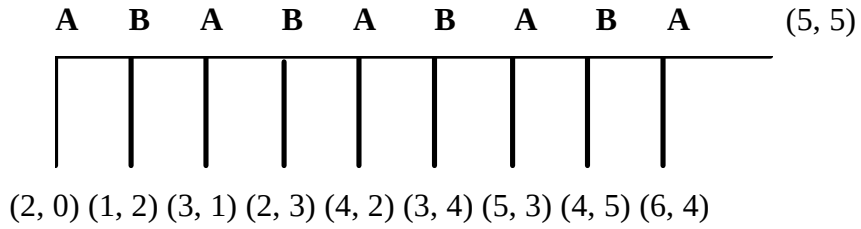
Figure 3.1.1 Centipede

In the Centipede agent **A** must choose between two options, forward (**F**) and down (**D**). In the game-tree in Figure 3.1.1 choosing **F** means moving forward in the right-hand direction and giving the opposite party an opportunity to make a choice of its own, whereas choosing **D** means pulling down and bringing the game to its conclusion without giving any further opportunity to the opposite party. (For technical reasons I was not able to place **F**'s and **D**'s in the figure.) Of course, the above game does not leave any room for **B** to give **A** a further opportunity of choice, but looking at the situation from **A**'s point of view the idea is clear enough for the present purposes. In any case, by choosing **D**, agent **A** secures an outcome (1,1), and the game ends, whereas by choosing **F**, agent **A** gives agent **B** an option between **F** and **D**. By choosing **D**, agent **B** achieves a better result (3) than by choosing **F** (2). That is, **D** is the admissible choice for **B** in that node. So, the dismantled, and game theoretically standard,

[76] This is what Bacharach (1992) calls the backward induction argument. Cf. also Pettit and Sugden (1989) and Sobel (1993).

account predicts that agent **B** chooses **D**, when a rational agent faces a situation as illustrated in Figure 3.1.1. Consequently, if rational, then, as the standard account predicts, agent **A** has no other option than to choose **D**, since in that case agent **A** would at least gain something (1) instead of nothing (0). That is, choosing **D** yields an admissible and unique solution to the problem of choice. So, both agents are losers as is 'we' in Hume's example. This presentation of the centipede corresponds to the game of the sequential **PD**: **D** is the dominating choice for agent **B**, hence, by backward induction, also to agent **A**, although both choosing **F** would lead to a "collectively" more acceptable result.[77] Because by following the requirements of rationality the agents end up with a collectively unsatisfactory result, some authors have seen a certain counter-intuitivity in this result. However, since choosing against the standard requirements is incoherent under the given circumstances and, thus, irrational, the intuitively more acceptable solution is ruled out - hence the backward induction paradox.

The Centipede becomes more easily understood when observed as a case with more options to be considered. In addition the nature of the paradox as involving backward induction becomes clear when the sequence of the game is extended. Let us consider the following sequential case:



Here the counterintuitivity of the standard account is more clearly manifested. The rationality requirement of the standard account dictates that the agent chooses **D** (and 2) instead of **F** (and possibly 5). If the sequence is stretched even further, let's say up to the outcome of 5 million, the logic of the standard account still dictates choosing an option securing 2 pennies instead of trying to pursue the 5 million. The back-folding deliberators seem to have no other option than to pull down. The standard account, applying the admissibility principle and the uniqueness principle together with the backward induction argument (**BI**), predicts that the players will deviate from a cooperative path (**F**), as is the case in the above example.[78]

The guiding logic a back-folding agent applies, depending on that agent’s capacity to make the

[77] Of course, the match between the games is not perfect, since in the **PD** it is rational to defect (play **D**) irrespective of the other's choice, whereas in the centipede it would pay for **A** to play **F**, given that **A** could somehow be assured that **B** plays **F**.
[78] Cf. Sobel (1995).

necessary recursions, goes as follows: (1) In a situation in which the players face a sequential game, it is a common belief that the players are rational.[79] Therefore, it is a common belief that they will be rational at round *n* and that therefore they will expect each other to act according to the requirements of admissibility and uniqueness. (2) It is a common belief at round *n - 1* that the players are rational at round *n*. Therefore, at round *n – 1*, the players will expect each other to act according to the requirements of admissibility and uniqueness at round *n*, and that, at round *n – 1*, they will expect each other to expect that they will act according to the requirements of admissibility and uniqueness at round *n*, and that (1). (3) It is a common belief at round *n - 2* that the players are rational at rounds *n - 1* through *n*. Therefore the players will expect each other at round *n - 2* to act according to the requirements of admissibility and uniqueness at rounds *n - 1* through *n*, and they will expect each other to expect that they will act according to the requirements of admissibility and uniqueness at rounds *n - 1* through *n*; and that (2). This process continues up to a certain point depending on the capacity of making the necessary recursions and the length of the sequence. Now, when the agent faces a situation such as the sequential **PD** or the Centipede, the dismantled analysis predicts that the agent in a sense deviates from the evident cooperative path.[80] So, the ever so reliable mechanism produces an unsatisfactory result, since the alternative option is ruled out due to incoherence.

4.1.2 Making a Mistake

In contrast to the forecast provided by the dismantled standard account, satisfactory solutions to the problems under discussion are repeatedly encountered. Could it then be possible that rational agents occasionally made mistakes that explain deviations from the expected solutions? Binmore (1987, 1988), following Lewis (1976), plays with the idea of the counterfactuals of the sort: "if kangaroos had no tails, they would topple over". Binmore asks what would happen if a perfectly rational agent made a sequence of irrational moves. Putting things this way means furnishing the account with liability to errors on the part of the agents and giving up the assumption that we are dealing with ideal agents. The kangaroos that have no tails are not ideal kangaroos, and they are, thus, liable to topple over

[79] 'I am rational' means that I consistently and coherently infer from my current beliefs what the best course of action is to satisfy my desires. From my experience I have learned that the principles of admissibility and uniqueness provide a reliable mechanism to this end.

[80] As a matter of fact, the players do not under the given circumstances deviate from any cooperative path, since such a path is unavailable to them. It would be more correct to say that given the necessary recursions and assuming that the standard requirements hold, a rational agent is expected to choose **D**. Since the account is supposed to be reduced to its hard core, no other option cannot be considered rational. This is due to the idealization of the model. In the dismantled account no mistakes take place, and the players never reach the boundaries of their belief systems.

occasionally. That is, if we want to continue the game started by Binmore, the kangaroos under discussion are liable to err because they are incomplete, i.e. they have no tails. More importantly, furnishing the agents with the liability to err saves mutual belief in rationality.[81] In the case in which the choice of an agent is apparently inconsistent with what the standard requirements recommend, rationality of the agents is appreciated in presence of incoherent behavior by explaining the apparent irrationality to be a product of the liability to err. In other words, the background theory is changed by allowing errors in the account, and these errors do not come without altering the point of view of the agents in the circumstances they are. For instance, in case of the sequential **PD** or the Centipede, this means that when an agent finds herself on a cooperative path, she is not necessarily warranted in giving up the assumption of the mutual belief in rationality.[82] Being mistaken is not a symptom of irrationality. It is a symptom of incompleteness, and, as Binmore puts it (1987, p. 198), in essence, all rationality is incomplete.

However, this account, call it the mistake account, does not have much bearing given that we are dealing with otherwise ideal agents that are incomplete merely with respect of being liable to stochastic errors, such as trembles. Stochastic errors do not carry e.g. the recurrent cooperative pattern that guarantees satisfactory solutions in the given circumstances, since in the name of consistency the mistaken behavior should be corrected when it is spotted, given that the agent follows the reliable mechanism provided by the requirements of admissibility and uniqueness. Although a mistake is not necessarily a symptom of irrationality, a conscious mistake does not survive from being that, as it falls into a class of self-deception. To recall, the requirement of admissibility states that no one knowingly chooses a dominated option, and making a conscious mistake is exactly what knowingly choosing a dominated option is about.

Trembles and stochastic errors in general do not refer to errors inside the belief systems of agents. In contrast to trembles mistakes due to incomplete belief systems may have bearing in explaining cooperative patterns in the studied circumstances for the reason of yielding a systematic error in the deliberation process. For instance, blindspots may provide the required explanation.[83] However, blindspots are not available as the explanatory basis when the agents are supposed to be ideal in all other respects except for their the liability to stochastic errors. The account must be supplied with boundaries of the belief system in order to warrant the required kind of incompleteness while still asserting the mutual beliefs in rationality in the presence of non-equilibrium behavior. To the extent that

[81] However, this being the case we cannot be dealing with ideal or perfect rationality, but incomplete rationality.
[82] Cf. however, an alternative argument provided by Bicchieri (1990, 1993) and the discussion in Chapter 5 in this essay.
[83] Cf. Sorensen (1988) for a detailed discussion.

the deviations from the standard requirements are treated as mistakes due to external or internal causes, the present analysis - the mistake account - is in accordance with standard game theoretical analysis, except, of course, for the matter of attributing rationality to the agents despite their being occasionally mistaken.

Even if the central implications of the boundaries are left aside for a moment, it is obvious that an account furnished with an incomplete belief system, such as a belief system containing blindspots, suffices for sustaining e.g. a prevailing pattern of cooperation. Although such a pattern may be expected to dissolve in the process of completing the belief system in the case of otherwise ideal agents - that is, ideal with respect of all cited features expect that of being liable to err. Thus, an account, in which the agents are still supposed to be ideal with respect to their capacities of making the necessary recursions long enough to fill up the incompletenesses due to e.g. blindspots, may still sustain occasional cooperative patterns in behavior.

Still, the resilience of rationality states that when a mistake is spotted, it is corrected in the light of coherence and consistency. Accordingly, coherence and consistency are seen as dispositional or structuring features that are activated when necessary. For instance, when a rational agent discovers a blindspot in his belief system, or when he faces a situation in which reconsideration of e.g. the other agents' expectations becomes current, a rational agent tends to correct his beliefs according to the habits of inference, e.g. according to the requirements of admissibility and uniqueness.

Things are not, however, as simple as the above considerations may imply. Allowing the liability to make mistakes in the belief system, we introduce incompleteness in the sense of boundaries to the depth of beliefs as well as in the "natural" capacity to make the necessary recursions in order to arrive at the correct choice. Things being thus, it may happen that a fault in one's belief system goes unnoticed and a cooperative pattern may survive virtually infinitely. But despite this possibility the main question remains. Provided that the emergence of cooperation, and of society in general, is a matter of resolving the **PD**,[84] the mistake account suggests that the cooperative solution be ultimately based either on mistaken reasoning or irrationality.[85] As cooperation is witnessed in our everyday life, and it does not seem right to think that it is only due to a mistake or e.g. a self-deception that cooperation succeeds, the notion of rationality calls for re-assessment. However, it should still be kept in mind that the mistake account provides a minimal requirement for the mutual belief in rationality of not collapsing when a cooperative choice is witnessed. So, making a mistake may occasionally explain a certain, apparently incoherent action performed by rational agents.

[84] To support this claim cf. e.g. Hardin (1982) and Taylor (1978, 1987).

4.1.3 Ordinary Rational Agents

Rational expectations become, in a sense, fuller when the boundaries in the depth of beliefs are given a wider interpretation than merely liability to err. The main implication of the boundaries is a possibility of inducing rationality to take non-equilibrium paths, such as cooperation in sequential **PD**'s. Furthermore, this inducement is not necessarily a product of a mistake in calculations, but a product of trying to signal one's willingness to adopt a more cooperative path that could lead to a more satisfactory result in the long run than the dismantled analysis allows. This result is supported by several authors who have claimed that it is intuitively suspect to hold that in e.g. sequential **PD**'s rational agents would defect from the first round, since permanent defection does not lead to the best results.[86] Nor can defection at round 1 be shown to be a uniquely rational strategy when boundaries are taken for granted, given that the sequence goes beyond the boundaries. Then, defection is not fully rational from the folk point of view.[87] In other words, in asserting fallibilism and boundaries in the depth of beliefs and recursive capacities, an important step toward folk rationality is taken.

Let us look at the argument in which the ordinary agents face a sequence that goes beyond the boundaries. The aim of the argument is to support the claim that as far as we are dealing with imperfect finite agents, the common knowledge of rationality does *not* necessarily break down e.g. in the act of cooperation.[88] The following scheme is assumed to provide a mechanism for boundedly rational agents to induce rationality to paths that are ultimately underdetermined. Call this scheme of the mechanism of inducing rationality the (**IR**): (i) the cooperative act must provide a credible signal for further cooperation on part of the agent; (ii) the cooperative act must in part be based on a credible signal for further cooperation on part of the other agent(s); and (iii) (i) and (ii) are commonly known.

This scheme can be given the following interpretation. Suppose there are two agents, **A** and **B**. For the sake of clarity, **A** is she and **B** is he. Suppose that agent **A** cooperates in an arbitrary round (early enough) under circumstances in which the sequence of possible interaction situations exceeds the agents' normal capacity of making the necessary recursions.[89] Agent **B** may then think that **A** believes

---

[85] Say, self-deception, wishful thinking or *akrasia*.
[86] Cf. Pettit & Sugden (1989) for a more detailed discussion. The most acknowledged single case on the issue is the TIT-FOR-TAT introduced by Robert Axelrod. Cf. e.g. Axelrod (1981).
[87] Cf. also the discussion in Rescher (1988): rationality essentially aims at giving the best reasons.
[88] Cf. Bicchieri (1989) and Pettit & Sugden (1989) for discussion.
[89] I believe with Anderlini (1995) and others that it is possible to stretch the recursions longer than the limit. But given the ordinary capacities and the requirements of the given situation, the costs of making any further recursions are greater

that by cooperating in that round she can induce **B** to cooperate in the next or immediately following rounds. The initial cooperative act of agent **A** may be considered rational if it can be held that she can induce **B** to believe that his cooperative act in the forthcoming round(s) could lead **A** to believe that **B** believes that by cooperating in a round in question he can induce **A** to cooperate on at least one of the rounds that follows **B**'s corresponding choice, e.g. at the next-to-the-next round. Agent **B**'s corresponding act of cooperation (the act induced by **A**'s initial act of cooperation) may be considered rational if it can be held that the initial cooperative act by **A** has in part given **B** a reason to believe that **A** will continue cooperating provided that **B**'s corresponding act of cooperation is in part based on **A**'s act of cooperation, and that **B** believes that by cooperating in that round he can induce **A** to believe that it is rational for her to cooperate in the following rounds, e.g. at the next-to-the-next round, as well. Furthermore, **A**'s corresponding belief in the rationality of cooperation (induced by **B**'s act of cooperation that was induced by **A**'s act of cooperation) may be considered rational if it can be held that it is in part based on **B**'s previous act of cooperation, which is in part based on **B**'s belief that he can induce **A** to believe that her (**A**'s) following act of cooperation can induce **B** to believe that **A**'s act of cooperation can induce **B** to believe that his (**B**'s) act of cooperation can induce **A**'s further cooperation, and, furthermore, that **A** believes that by continuing cooperation she can induce **B**'s further cooperation! This system of intertwined beliefs in the rationality of the corresponding choices of action, e.g. cooperation, maintains the expectations of rationality of the prevailing pattern as long as the corresponding choices are not strictly dominated by any other choice. Under the observed circumstances an act of cooperation cannot be shown to be dominated by any other strategy, while it still can be said to yield a satisfactory solution to the problem at hand.[90] Intertwined expectations of the rationality of the prevailing pattern may survive until the sequence of the game is manageable within the ordinary capacity of the agents, in which case the prevailing pattern may dissolve.

The main point of the mechanism of inducing rationality, the (**IR**), is that it shows that the players are *capable* of inducing (by reciprocations) each other's rational expectations, e.g. regarding cooperation, under circumstances in which the sequence exceeds the capacity of making the necessary recursions. That is, the agents are capable of making the option under discussion *rationally conceivable*. Especially, this means that agents may expect rational behavior from each other even when the completion of the deliberation process (in the sense of satisfying the conditions of admissibility and uniqueness) is virtually impossible. Furthermore, inducing rational expectations in this way is not

---

than the benefits, or they are expected to be greater, thus making further recursions possibly irrational. Economists would speak about diminishing marginal returns.

[90] I shall argue for this claim in more detail in Chapter 5 of this essay. Cf. e.g. the assurance problem.

necessarily the result of a mistake; instead the inducement takes place by consciously signaling the option that under the circumstances seems rational. Inducing rationality of cooperative option may be rational even if the agents are aware that the end of the game is relatively near. If an agent succeeds in inducing further cooperation by making a cooperative move, the players may have gained something when the remaining sequence no longer goes beyond their recursive capacities and the standard requirements apply. That is, given the benefits of the cooperative pattern, the harm of being the second one to "pull down" is not so great as to prevent conformity to cooperation.

This is true unless, of course, the sequence of the game is not sufficiently long to exceed the boundaries that are applied in everyday deliberations, in which case the standard requirements (of admissibility and uniqueness) are fully applicable. In the case of a sufficiently long sequence, the standard requirements lose their applicability, since the necessary recursions go beyond the ordinary capacities of the players. The players also make a forecast about the following sequence (the remaining part of the game) according to these capacities, and constantly update it according to their experiences until the standard requirements bite again. In the case of sequential **PD**, for example, the remaining sequence may be too short to try to induce the other player(s) to cooperate, as the prospects of cooperation may not seem favorable, or the inducement appears to be irrational. At this point the standard requirements of admissibility and uniqueness regain their force. In other words, rationality in terms of the standard requirements is, in a sense, dispositional, and it is triggered when circumstances are right. That is, when the end of a finite game is close enough, the inducement of the rationality of e.g. cooperation can no longer be maintained. This, in turn means that at some point close to the final round (round *n*) it may be expected that cooperation will decline depending on the capacities of the players, sooner or later.[91]

4.2 Folk Rationality: a Basis for Social Habits

Instead of appealing merely to mistakes, the (**IR**) proposes a two-way mechanism for appreciating the rationality of the agents that deviate from the perfect-equilibrium path recommended by the standard requirements. Especially, the (**IR**) proposes certain specifications to the notion of rationality embraced by the standard account. The alternative account proposed by the (**IR**) - *the signal account* - operates with the idea that the agents update their forecast on the basis of the signals emitted from the other agents' behavior, and they acknowledge the signals by making the corresponding choices. By making

the corresponding choices the agents signal their willingness to maintain the proposed pattern. The two-way mechanism looks back on past events seeking further reasons for making the choice. On the other hand, the mechanism looks ahead to what can be expected by inducing further expectations on the basis of the present choice, that is, a signal to the other agents about the pattern that one is willing to reinforce under the circumstances.

The signal account holds, thus, that under certain circumstances, such as sequential games, deviations from equilibrium paths are not treated as mistakes that lead to intentional behavior or not, but as signals about how one is going to play the game.[92] The idea is that an agent gives signals about her profile by making certain choices, and these signals force the other players to reconsider the strategy profile and possible moves of the agent. The difference between the mistake account and the signal account is not just a matter of figure of speech. The signal account appreciates the rationality of an agent's belief system prior to her choices. Even if a choice is unexpected, this does not justify branding it as irrational. Instead, as already argued, it provides justification to reconsider the strategy profile of an agent, but without suspecting her rationality. For instance, in the signal account a player that chooses cooperation in the first round of a sequential **PD** is treated as rational, since the cooperative move may be considered as an intentional signal (as opposed to a mistake - such as a tremble or a blindspot) that the player is willing to cooperate. The mistake account, on the other hand, allows the deviations from the equilibrium path only by mistake. Mistakes as such do not make the agent irrational, for it is not irrational to make mistakes, but when mistakes are spotted they should be corrected, even if the results are unsatisfactory. The 'should' here refers to a general desire, discussed in the previous chapter, that the belief system corresponds to the real world. The account being inductive stresses the point that e.g. blindspots do make a difference, since inductive inferences may find different paths when blindspots are filled up. This is in fact the case with the sequential **PD** in the standard account: eliminated mistakes lead to noncooperation in the first round.

Thus, to return to the issue, the starting-points of the signal account are (i) to appreciate the rationality of other players, even if their moves appear to be inconsistent with the standard requirements, and then, (ii) given the information from their current experiences to make their inferences as to the best reply to a certain expected strategy profile.

The signal account proposes specifications to the requirements of rationality in order to make the analysis intuitively more appealing than the standard account does, since it appreciates the players'

---

[91] As Sobel (1995) claims, if the sequence is short (that is, it does not exceed the capacity of the ordinary finite players) the players may be expected to play according to the standard requirements.
[92] Cf. Bicchieri & Schulte (1997).

choices as rational, although the players may deviate from the equilibrium path recommended by the standard account.[93] As we are willing to appreciate the standard requirements as an expression of the core notion of rationality, these deviations must be accounted for while still appreciating the rationality of the deviant agents. This being the case the need for constraints as to what may be considered as rational, if the requirements of the standard account are considered strictly and narrowly, becomes obvious. That is, certain rules for the signals are necessary.

In the game theoretical literature the signal account is most clearly expressed in the ideas of forward induction (**FI**), correlated equilibria and rationalizability.[94] All these accounts suggest certain specifications to the standard solution concepts, and, thus, to the attitudes toward rationality. Still, the idea of maximizing the expected utility is appreciated in these accounts. They are all in accordance with the (**IR**) as well. Al-Najjar (1995, p. 174) has stated that the principle of **FI** suggests that belief formation is a dynamic process that is inextricably tied to the agents' past behavior in inducing the expectations of other agents instead of applying any particular equilibrium that the agents play. That is, the two-way mechanism of the (**IR**) is inherent in the **FI**.

Aumann (1987) states the case of correlated equilibrium a bit differently, but it also can be interpreted as being a special case of the (**IR**). According to Aumann the agents may achieve a correlated equilibrium if they can tie their beliefs to an external event space. If this external event space is to be considered non-artificial, it must accord with the (**IR**). This means, especially, that the event spaces the beliefs are tied to are the same in their relevant parts and that the action choices of the corresponding agents are non-independent. One way of guaranteeing this is to permit agents to induce expectations in each other by making the choices they make and, on the other hand, to permit them to base their choices in part on past events that consist of the choices made by the agents.

Rationalizability, as introduced by Bernheim (1984) and Pearce (1984), formalizes the idea of reasonable choices in contrast to referring to (perfect) equilibria. Bernheim (p. 1011) proceeds on the basis of the following assumptions: (1) The choices of the agents cannot be determined from the basic information, (2) the choices of the agents are based on their subjective probability distributions over the possible events, and (3) assumptions (1) and (2) are commonly believed by the agents. Thus, rationalizability states that there may not be a unique solution to the problem of choice. Furthermore, the principle of rationalizability leaves open the option of inducing expectations to other agents, since under the circumstances of incomplete knowledge and in view of assumption (2) the agents may update

[93] Note that the deviation takes place given the strict interpretation of the admissibility and uniqueness criteria.
[94] Cf. Ponssard (1990), Al-Najjar (1995), Aumann (1987), Vanderschraaf (1995), Bernheim (1984), and Pearce (1984) for discussion of the mentioned topics.

their information about the situation according to the moves of the opposite parties, and this is commonly believed by everyone. Furthermore, in view of assumption (1) it is not irrational to try to induce the expectations, and this is commonly believed by everyone. Of course, these beliefs must be interpreted *de re*. Thus, the principle of rationalizability is also in accordance with the (**IR**), and it, consequently, represent an example of the signal account just as the **FI** and the correlated equilibrium analysis do.

In the signal account it is no longer required that the solutions be self-enforcing in the sense of the standard account. Instead, when pursuing an eligible solution, certain additional factors, such as precedents, external events and further (rationalizable) reasons, are referred to. Especially, seeing the issue in terms of e.g. a path may clarify the concept of rationality applied by the signal account. However, this does not mean that the standard requirements of admissibility and uniqueness, as well as the **BI**-argument, should be rejected. After all, they state the core conception of rationality. To embrace this conception, it must be noted that all the examples of the signal account and the (**IR**) coincide with the standard account with incomplete information. That is, when the boundaries are taken into account, the underlying principles of rationality are essentially the same in both accounts. Assuming bounded rationality, however, makes the account more flexible due to its dynamic characteristics than the account that refers merely to incomplete information.

The signal account, however, is not unproblematic. A main problem of providing a resolution to the problem of unsatisfactoriness in terms of the signal account, and e.g. the (**IR**) is that it introduces a problem of indeterminacy. If an alternative is offered to an unintuitive but sound result, there is no longer a unique solution to the problem of choice. This is an important issue that I turn into in the next chapter. It suffices to say here that because folk rationality is essentially incomplete there is room for social habits to flourish. In fact, because of its incompleteness folk rationality seems to require the support of social habits, for they promise a resolution to the problem of indeterminacy in terms of rationality, although without the present deliberation of choice. That is, the need to explicate collective rationality becomes relevant in part because individual incomplete rationality cannot handle the problem of indeterminacy. Before proceeding to the issue of social habits, a couple of minor problems from the point of view of the present essay are worth mentioning: the problem of providing inconsistent recommendations and the starting-up problem.

4.2.1 Inconsistent Recommendations

The immediate problem with the signal account is its apparent inconsistency with the standard account,

since the recommendations of the accounts appear to be incompatible with each other. This appearance is not far-fetched, since, after all, one of the basic aims of the signal account is to solve the problems that the standard account has been incapable of solving.

The aim of the signal account, and consequently of the (**IR**), can be connected to e.g. the discussion of the resolution of the backward induction paradox. Pettit and Sugden (1989, p. 171) following Frank Jackson (1987) anticipate the signal account. They claim that the backward induction paradox can be resolved by recognizing that rational players under the given circumstances are not in a position to run the backward induction argument. If they were, then, necessarily, the players could not believe that the common belief in rationality would survive whatever moves the players make.[95] An act of cooperation would cause the common belief in rationality to collapse. This result seems unavoidable to the extent that we are dealing with ideal agents; however when it comes to agents who are prone to make mistakes, or to boundedly rational agents in general, this is not so. This is the main reason for looking for an alternative resolution to the backward induction paradox from the point of view of bounded rationality. In that case the common belief in rationality may survive irrespective of the choices of the players. Assuming that the players are liable to make mistakes, an act of cooperation does not warrant condemnation to irrationality, and, thus, to the breakdown of the common belief in rationality. Assuming that the boundaries in the depths of the beliefs are relatively low, it is far from necessary that the players cannot have confidence in the survival of the common belief in rationality in the case of an act of cooperation. Pettit & Sugden's argument anticipates the present account, since it indicates the essential incompleteness of rationality by stating that the agents who face a finite sequential **PD** are not necessarily in a position to run the **BI**, which would not be true of the perfectly rational ideal agents as postulated by the classical game theoretical approach. Therefore, under the circumstances of bounded rationality, besides having to apply the **BI**-argument, other options for making the choice are available for the agents, as well. This means that the uniqueness of the solution cannot be guaranteed,[96] which in turn opens the gate to the apparent inconsistency between the recommendations of the **FI** and the **BI**.

The (**IR**), however, provides a resolution to the apparent problem of inconsistency. If and when the sequence of the game is short enough, the (**IR**) no longer supports the inducement of rational expectations regarding the actually non-admissible choice,[97] such as an act of cooperation in a finite

[95] Cf. Pettit and Sugden (1989), pp. 172-174 for a detailed argument.
[96] Cf. e.g. Bernheim (1984), Pearce (1984), Ponssard (1990), Bicchieri and Schulte (1997) for arguments on the redundancy of the uniqueness.
[97] By the actually non-admissible choice I mean a choice that appears to be non-admissible under the circumstances of perfect information. This applies also to bounded rationality when the agents are capable of completing their recursions of beliefs in the forthcoming sequence.

sequential **PD**. Under circumstances in which the sequence is short enough, the rules of finite games with perfect information may be expected to apply, and hence the players are in a position to run the **BI**. On the other hand, when the end of the game is beyond the boundaries, the (**IR**) permits the induction of rational expectations about the choices that under the circumstances appear to be (at least weakly) admissible. That is, there is no other option that would strictly rule out the induced option, e.g. an act of cooperation. If an act of cooperation can be expected to induce further cooperation, it is (weakly) admissible. For instance, in the case of a finite sequential **PD**, cooperation can be expected to be admissible as long as the sequence goes beyond the boundaries of the agents' ordinary recursive capacities.[98] Thus, the (**IR**) yields a principle that can be applied under circumstances of both bounded and perfect rationality,[99] and it explains the apparent inconsistency between the accounts in terms of the depth of beliefs and beliefs in the depth of beliefs of the agents.

4.2.2 The Starting-up Problem

The (**IR**) as a two-way mechanism also confronts an instant problem concerning the first-round choice. There is no pattern to reinforce, nor is there any relevant information on the other agents' capacity to make the necessary recursions. Furthermore, there are no past events that one could use in order to update the present beliefs and expectations. It is usually required that the agents have a common history long enough when dealing with corresponding accounts, say the correlated equilibrium analysis by Aumann (e.g. 1987) or the forward-induction analysis (by e.g. Al-Najjar (1995)). This problem does not undermine the present account, though, since it is flexible enough to deal with this incompleteness. After all, rationality is essentially incomplete and rational choices are essentially underdetermined when boundaries are taken for granted.[100]

The starting-up problem can be resolved by showing that an inference of a cooperative move is conceivable even if the cooperative precedents are missing. For this purpose a principle of conceivability can be applied. The idea is to make coordination available and a cooperative move rationally conceivable for the agents, in contrast to determining cooperative moves or making cooperation in the present situation necessary. An option to defect is still available, but the agents may take the

[98] Cf. Bacharach (1992) and Sobel (1993) for accounts of ordinary finite agents and a discussion of the depth of beliefs.
[99] Cf. also Bicchieri and Schulte (1997) p. 301. They present the concept of IWD (iterated elimination of weakly dominated strategies) in order to resolve the inconsistency problem between FI and BI. They claim that FI and BI both follow from this principle. Also Al-Najjar (1995) proposes a novel solution concept, the forward induction equilibrium (FIE), that is claimed to provide the required constraints to the FI in order to avoid the inconsistency with the BI.
[100] Cf. e.g. Binmore (1987), p. 198.

cooperative move under consideration.

One way of rationalizing the cooperative option is to apply a fictitious play on the part of agent **A**. As **A** can rationalize the first-round cooperation by believing that she can induce **B** to cooperate by inducing **B** to believe that he, in turn, can induce **A** to further cooperation, so can **A**, by applying the very same mechanism, place herself in her beliefs into the position of **B** and make a fictitious play from the point of view of **B** if and when **B** was given the opportunity to make a choice after **A**'s cooperative act. That is, **A** may use a fictitious play to give a further reason for rationalizing the first-round choice that is normally rationalized by appealing to past events. The trick is that agent **A** makes the choice as if it were in part based on experience. But there is also a more serious point to be made. By means of fictitious games an agent may test the *degree of assurance* concerning the cooperative option.[101]

The basic idea is that once a cooperative move in a sequential situation is witnessed, a further reason for cooperation, besides mere conceivability, has emerged, and all the witnessed acts of cooperation and coordination of actions in general add up to general conformity to the present pattern of cooperative moves. That is, the idea of the principle of conceivability is to give reasons for making a cooperative option available for the agents in the sequential situations, and once a cooperative move is witnessed, a basis for cooperative precedents is created. There is no need to determine cooperation. Making it available suffices, and once available the rest may happen by a snowball effect of inferring further expectations on cooperation, provided, of course that the first round of cooperation ever happens. However, when the first-round cooperation is made rationally conceivable, a problem of indeterminacy enters the scene.

5 SOCIAL HABITS AS EXPRESSIONS OF COLLECTIVE RATIONALITY

5.1 Facing the Problem of Indeterminate Solutions

The signal account has been widely associated with social habits and, consequently, with norms, institutions and conventions.[102] This is due to the fact that the signal account as presented in the

[101] Cf. Chapter 5 for the discussion of the assurance problem.
[102] Cf. Aumann (1987) and Vanderschraaf (1995a), who explicitly link the correlated equilibria account with Lewis' (1969) analysis of social conventions. Bernheim's (1984) and Pearce's (1984) notion of rationalizability is essentially

previous chapter introduces the problem of indeterminacy, which, in turn, undermines the predictive power of the model if agents do not take previous events into account. The general problem of indeterminacy of solutions in the models of repeated and sequential games[103] points directly to the problems of coordination which Lewis (1969) hoped to resolve by means of social conventions.[104]

Roughly put, the problem of indeterminacy means that an agent has no *a priori* means to correctly predict an opponent's moves in a given interaction situation. More generally, the problem of indeterminacy represents a serious challenge to attempts to account for rational behavior in many social interaction situations in which apparently rational behavior is witnessed every day. That is, the standard notion of individual rationality appears to be unable to explicate certain evidently rational actions. One way for accounting for rational behavior under these circumstances stems from the signal account, which in part rests on an *a posteriori* way of inferring expectations of rationality. That means, ultimately, that the agents base their moves (in part) on conformity to a prevailing pattern rather than presently deliberating all the relevant alternatives. In order to deal with a situation in which agents recurrently confront the problem of indeterminacy, the agents tend to adopt and follow social habits, e.g. conventions, norms and institutions, and thereby save their inferential capacities for other occasions.

Besides the problems of coordination and, consequently, the problem of indeterminacy, social conventions, norms and institutions are claimed to resolve the problems of unsatisfactory outcomes in social interaction situations, such as the **PD**, as well.[105] Note, however, that the **PD**s fall under the class of indeterminacy problems when iterated - that is, sometimes a unique solution can no longer be found when the problem is stretched into sequences. In open-ended **PD**s and other sequential problems of the

---

based on the idea of common knowledge, introduced by Lewis. Brandenburger and Dekel (1987), on the other hand, associated the notions of rationalizability and correlated equilibria with each other. Cf. also Vanderchraaf (1995b), who pointed out this connection. Ponssard (1990), on the other hand, has explicitly linked the idea of forward induction and bounded rationality arguments to the emergence of social conventions. Cf. also Al-Najjar (1995). Ullmann-Margalit (1977) defends Lewis' approach and generalizes it to apply to social norms and institutions as well. Cf. also Schotter (1981). Cf. Tuomela (1995), for an alternative approach to conventions and norms.

[103]Cf. Harsanyi & Selten, (1988), on the general problem of indeterminacy. The difference between the repeated or iterated games and the sequential games must be emphasized: The former can be reduced to their normal form, such as e.g. a one-shot **PD** or **CG** or **BS** etc., whereas sequential games cannot easily be reduced in this fashion. The most natural way to illustrate the sequential games is the extensive form.

[104] I have been criticized for discussing the findings of mathematical models on rationality and the real life situations in parallel when it comes to resolving certain problems in terms of social habits. However, there is a substantial literature on the topic in which mathematical models and real life games intersect. For instance, Aumann (1987), Ponssard (1990) and Vanderschraaf (1995b) are directly dependent upon Lewis (1969) and try to construct mathematical models of situations which he described. On the other hand, Lewis bases on the discussion of these mathematical models, e.g. Schelling (1960). Also Ullmann-Margalit (1977), Schotter (1981), Bicchieri (1990, 1993) and Vanderschraaf (1995a) apply game theoretical findings to discuss e.g. social norms. The same tendency can be found in all discussions of the signal account, e.g. Al-Najjar (1995) and Bicchieri & Schulte (1997), in that they try to resolve the game theoretical problems with the constraint of respecting common intuitions about certain outcomes.

type, such as the Centipede, there are many cooperative options that compete with the standard perfect equilibrium option.[106] Hence, under dynamic circumstances, social conventions, norms and institutions are able to resolve the indeterminacy problem of the **PD**-type situations, as well. The finite iteration of the **PD** in case of ordinary agents approximates the situation that ideally rationally agents face in the infinitely iterated game.[107] This means that cooperation can be rationally conceivable even when the rounds of the game are limited. Consequently, the perfect equilibrium solution and the rationally conceivable cooperative solution in coexistence introduce a problem of indeterminacy. The distinguishing feature of **PD**-type situations from e.g. the pure coordination problems under dynamic circumstances is that in the former the cooperative pattern tends to disappear close to the end of a finite interaction situation, whereas in the latter the cooperative pattern is viable even on the last round.[108]

Taken literally, however, Lewis' analysis applies merely to pure coordination problems. By all means, the study of the coordination problems is an instructive way to approach the issue at hand. A basic example of a pure coordination problem is e.g. a meeting-point game. Suppose there are two commonly known meeting spots, say **C** and **D**, in the town and two persons wish to meet each other, but they have forgotten which one was the place they were to meet. Both of the persons in the given situation must decide whether to go to **C** or **D**, each according to his expectation of the other's choice, in order to choose **C** if and only if the other chooses **C**, as well. Described as above, the game represents a game of pure coordination. Even though it seems simple, the resolution to the problem is not trivial, since it can be solved only if both parties can find a way to coordinate their activities. In a strategic form (= normal form) the example is presented as follows:

	C	**D**
C	(1,1)	(0,0)
D	(0,0)	(1,1)

Figure 4.1.1 the game of pure coordination in the normal form

The game has two pure-strategy equilibrium-points, which causes a problem how to determine a particular action. Lewis suggests (p. 25) that the players are more likely to succeed if their mutual

[105] To begin with, cf. Ullmann-Margalit (1977) and Schotter (1981).
[106] Cf. the lively discussion on the topic beginning from e.g. Axelrod (1980).
[107] Cf. e.g. Bicchieri and Schulte (1997) for support for this claim.
[108] This result is due to the backward induction argument presented in Chapter 4. The backward induction argument holds if the world is small enough. Cf. Savage (1972) and Binmore (1993) for discussion.

expectations are suitably concordant. Suppose e.g. that you and I face a situation similar to the example case, then I may go to **C**, because I expect you to, while you do, because you expect me to. In general, Lewis states (p. 25), each may do his part to reach one of the possible coordination equilibria because he expects the others to do theirs, thereby reaching that equilibrium.

The resolution of such problems by means of social habits may, however, take place only under certain type of dynamic circumstances. The agents may resolve the corresponding problems by means of the cited social habits if a sufficient amount of coordination between the agents has already been witnessed. That is, the resolution depends not only on the common knowledge (or, mutual belief) of rationality, but on the common knowledge of sufficiently coordinated past events as well. Without the element of recurrence the discussion on (social) habits would be meaningless. This is an important constraint of the present account. Without a sufficient amount of coordination the physical basis for conformity would be missing. Without a sufficient amount of coordination the agents would have no evidence of being capable of inducing (by reciprocation) rational expectations e.g. about cooperation and future conformity. Consequently, the agents could not trust their present moves of conformity to any deliberational rational basis that would justify it.[109] However, this rational basis may not have been articulated in any way. This can be said to be true especially when it is a case of the present move of conforming to a social habit. Without the support of sufficient coordination in the past, the present move would be merely a rational choice in contrast to conformity to a certain pattern that social conventions, norms and institutions evidently create. In short, resolution by means of social habits of this type requires that there be an existing path trodden by the agents, and once such a path has been trodden, the appropriate expectations may apply.

The signal account, and the (**IR**) presented in the previous chapter, makes the emergence of cooperation, and consequently the gradual accumulation of coordination, conceivable, but it still does not seem to suffice to account for full-fledged conformity. After all, a move recommended by the reliable mechanism, such as the (**IR**), is still based in part on present deliberation, but conformity to a certain pattern, such as a social norm, *typically* takes place without such reconsideration or present deliberation. Still, the (**IR**) is an essential part in justifying the social habit, although it does so indirectly. In conforming to a habit the (**IR**) works in the background making the habit at stake rationally conceivable, although it is not appealed to in making a conforming move.

To deal with situations in which the moves of the agents are seen as conforming to a certain pattern, rather than making a deliberated choice between two or more options, I introduce the idea of

[109] E.g. Gilbert's (1990) criticism supports this claim.

funk games. Funk games are games typically played without present deliberation or reconsideration about the move to be made at a given situation. They are games that operate with such notions as *conformity* to habits, conventions, norms, institutions etc. in contrast to the regular notion of choice familiar from the eductive game theory. Conforming to a social convention in the sense of Lewis (1969) expresses the idea clearly.

Despite their apparent incompatibility with the standard account of game theory, funk games are inextricably anchored to their eductive basis. Without a deliberational rock bottom, no funk game would emerge, although funk games do not seem to be reducible without a residual to the standard eductive analysis, either. Especially, funk games contain information of the common knowledge of rationality and coordinated past events in conformity to the corresponding path. This information comes in the generalized form of e.g. a social norm, an institution or a convention. Generally speaking, too much is required if ordinary agents must in each round deliberate this information over and over again, when they can simply conform to a previously deliberated and functioning option. In a sense, funk games, then, deal with *conserved* rationality. Besides being in a sense conserved in the form of the social habits, rationality is also *customized* and *blinkered.* Social habits in a sense customize rationality in that present deliberation, e.g. in the sense of the (**IR**), is not needed, since the conformity to a commonly known and accepted regularity is enough to ensure the satisfactory result in the given community or culture. The customs of a given society with a given past are (*de re*) expected to ensure satisfactory results in social interaction without squandering the capacity to make inferences by deliberating the best option in a recurring situation over and over again.[110] In addition, collective rationality realized by social habits is blinkered in that the alternative options are not typically reconsidered when conforming to the prevailing pattern. Instead, a routine move is made when acting on a social reason that is indirectly supposed to maximize the expected utility.

Funk games are, in a sense, a product of bounded rationality in that they emerge as a resolution to the problems of deliberation under circumstances in which the recursive capacities of the agents do not suffice for completing the deliberation according to the standard requirements of rationality. They are games that arise under circumstances of recurring situations. I take this to mean situations in which the termination of the game goes well beyond the recursive horizon of boundedly rational agents. When such a funk game emerges, certain simple, higher-order rules of thumb about how to act in the corresponding social situations take over. They provide a solid basis for action - i.e. they generate action in part based on common knowledge. Furthermore, funk games are played without present

[110] It is important to make a distinction between optimal use and wasteful spending of the capacity.

deliberation of the situation. In addition, as in the case of social facts, higher-order funk games may nest in lower-order funk games, and they also provide the foundation for higher-order deliberational moves, as well.[111] In short, funk games are based on an idea of building up a *routine* that is in part based on previously deliberated choices and in reciprocal expectations and which are implemented by making a routine move in the defined circumstances.[112] These routine moves may, then, take place as if they were based on an "instinct" rather than rational deliberation, although in actual fact they rest on an eductive basis rather than e.g. on a "moral instinct".

The aspect of conformity and common beliefs about this conformity add more to the analysis than the standard account is able to provide - namely, the social reasons for action. However, these social reasons cannot survive unless the circumstances are right, and I take them to be right when it is feasible to induce reciprocal expectations with respect to collective rationality. That is, viable social reasons require the emergence of funk games. Consequently, funk games are true games of collective rationality.

It is instructive to begin the argument for funk games by presenting their main antecedents and their critiques. I shall study the approach from the point of view of the findings of the previous chapters. That is, the focus is on finite and fallible non-Bayesian agents that are capable of inducing rationality of certain moves by applying a simple recursive mechanism, such as the (**IR**). At the end of the chapter, I shall finally present the main argument for adopting the idea of funk games. A central initiator of the present approach is David Lewis' game theoretic analysis of conventions. After summarizing this analysis I shall explore a way to expand his approach.

5.2 Social Conventions: Accounting for Social Reasons

According to Lewis (1969), social conventions are regularities of behavior in recurrent situations that are conformed to because of mutually concordant expectations, and this concordance is gained by a relatively long history of precedents.

Lewis' conjecture can be given the following interpretation. Previous choices set constraints to options available now and in future. The past sequence of choices forms a path of options that has survived from sequential eliminations and, thus, it has on its part set constraints on the available moves.

[111] This feature is backed by Searle (1995).

[112] Consider for instance a game of chess. Besides following the rules of the game in a non-deliberational fashion, the players may apply certain models according to which they play the game in the (virtually) non-deliberational fashion for a while or they may deliberate the next move very carefully leaving the forthcoming moves aside for a moment. In both cases the course of the game goes beyond their deliberational capacities.

The sequential elimination of ineligible alternatives in a sense paves the way for behavior that becomes regular when repeatedly chosen. When a sufficient degree of concordance is gained due to the experience of precedent choices, people may begin to conform to the prevailing regularity, which has, thus, become a social convention.

Seeing matters this way opens a wider perspective on conventions than that provided by Lewis himself. This observation is confirmed by the insight that Lewis focuses on solutions that go in terms of equilibria[113] rather than in terms of a rationalizable path as is done in the above interpretation. Interpreting conventions in terms of the rationalizable path emphasizes the self-enforceable characteristics of conventions, as is the case with Lewis' own interpretation. However, my proposal avoids a great deal of the criticism that has been levelled against Lewis' account,[114] and it also responds to the remaining criticism, as will be seen later. At any rate it must be said that Lewis' approach anticipates the conception of social habits discussed in this essay. Lewis' characterization refers to circumstances in which the coordination of the players has been already reached to a remarkable extent in previous rounds of the game. Furthermore, especially, it refers to circumstances in which the agents face a recurring situation loaded with past experiences and future expectations, and all this is commonly known. Under such circumstances it hardly can be said that the behavior of an individual is totally independent of external events, as the classic accounts of the game theory claim.

To make some progress on this issue it is instructive at this point to take a look at Lewis' analysis. Note that Lewis' definition contains certain apparently ambivalent features: it refers both to the notion of equilibrium and to the notion of conformity. I hope to clarify the entailments of this feature of the definition in the course of the forthcoming discussion.

The definition for social convention goes as follows:

(SC) A regularity **R** in the behavior 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 is common knowledge in **P** that

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

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

(3) Everyone prefers to conform to **R** on the condition that the others do, since **S** is a coordination problem and uniform conformity to **R** is a coordination equilibrium in **S**. (Lewis 1969, p. 58)

[113] To be more exact, Lewis focuses on the coordination equilibria, but even then he focuses merely on equilibria, and not on a path that has lead to this or that solution.

[114] Cf. e.g. Gilbert (1990) and Miller (1990).

In Lewis' definition there are plenty of features that call for explication. The three main features that characterize behavior of agents when they follow a social convention are common knowledge, conformity to a regularity and rationality. Let us begin with the common knowledge part of the right hand side of the definition.

5.2.1 Common Knowledge and Truth

The requirement of common knowledge of conditions (1) - (3) and the requirement of truth of conditions (1) - (3) express the epistemic conditions of the social conventions. To be more specific, Lewis' definition requires that in order for a regularity in the recurrent social interaction situation to be called a convention there must be common knowledge about conditions (1) - (3). Besides being characterized by common knowledge of precedents, past conformity and preferences, these conditions can be said to manifest a modicum of rationality.[115] Especially, condition (3) can be said to indicate rationality on the part of the agents, since it implicitly refers to maximizing expected utility. Condition (1) and (2) support the expectations of rationality by referring to the evidence on the past behavior and by stating what the agents believe they are presently expected to do. So, besides having common knowledge of the circumstances, the agents can be said to possess common knowledge of rationality, as well - and, incidentally, higher-order expectations about each other's rationality and forthcoming sequences of the game.

In the game theoretical literature common knowledge has been formulated in various ways. In short, something is common knowledge among a group of agents if it is known to all, it is known to all that it is known to all, and so on *ad infinitum*. Now, since we are dealing with boundedly rational agents, there must be a limit to the recursions to the knowledge (or beliefs) of the agents.[116] I think it is safe to assume that the level of necessary recursions is relatively low, while the agents at each round update their information by applying a simple mechanism (that includes the necessary recursions). This too, I think, should be considered to be common knowledge among the corresponding agents under the studied circumstances. That is, it is common knowledge that the agents are boundedly rational.

A formal characterization of common knowledge can be stated as follows:

(**CK**) It is common knowledge that *something **X** is the case* for a group of agents (*i, j*) iff
(1) Each agent *i, j* knows that *something **X** is the case*,

[115] Cf. Lewis (1969), p. 57.

(2) Each agent *i* knows that each agent *j* knows that *something **X** is the case*, and each agent *j* knows that each agent *i* knows that *something **X** is the case*,
(3) Each agent *i* knows that each agent *j* knows that each agent *i* knows that *something **X** is the case*, and each agent *j* knows that each agent *i* knows that each agent *j* knows that *something **X** is the case*, and so on (*up to a certain point).*[117]

The requirement of common knowledge is crucial for coordination of actions, since without common knowledge no interdependent moves could take place. However, this is claimed to be too strong an assumption in general. In some opinions the postulate of common knowledge is an incoherent, troublesome, and even impossible assumption to make.[118] Note, however, that common knowledge does not entail perfect or complete knowledge. It refers only to the range of knowledge that is, roughly put, openly accessible to and relevant for the mutual decisions or conformity to the corresponding pattern. As a matter of fact, when agents are assumed to be boundedly rational, the cited problems of common knowledge dissolve. Common knowledge and bounded rationality make a perfect couple for the present purposes, as was made clear in the previous chapter.[119]

Lewis builds up the definition of common knowledge in terms of beliefs. This being the case, it is only a matter of taste whether we prefer to use the term common knowledge or mutual belief in Lewis' sense.[120] According to Lewis, "it is *common knowledge* in a population **P** that *something **X** is the case*" if and only if some state of affairs **A** holds such that:
(1) everyone in **P** has a reason to believe that **A** holds,

---

[116] Cf. the previous chapter for discussion.
[117] Some may find this characterization problematic with respect to the constraint that the necessary recursions are taken into account only up to a certain point. It could be said that this procedure does not yield common knowledge. However, the constraint is grounded in the light of the incomplete skills and capacity of the ordinary agents. Furthermore, this formulation resists the infinite regress of recursions that the ordinary agents find inconvenient. A similar idea is backed e.g. by Binmore (1987, 1988, and 1993). He speaks about stopping-rules and guessing algorithm, as well as the essential incompleteness of rationality. In a sense, these ideas speak for the finiteness of embedded knowledge as well. Another formulation could be put as follows: (1) aKp, (2) bKaKp, (3) aKbKaKp, and so on *up to a certain level*, where aKp means 'a knows that p', etc. The same conditions hold, naturally, for b as well.
[118] In the previous chapter I made this point clear while discussing the **BI**-paradox. However, cf. Bicchieri (1989) who claims that the postulate of common knowledge makes the theory of games inconsistent. Cf. also Pettit & Sugden (1989). Dufwenberg & Lindén (1996), on the other hand, claim that common knowledge is not the issue. The problem of inconsistency, Dufwenberg and Lindén argue, arises even with minimal beliefs, which is contrary to what Bicchieri claims.
[119] An alternative approach to dealing with common knowledge and belief by an infinite conjunction etc. is the fixed point approach. In it the syntactical infinity of beliefs is cut short by means of a formula in which the notion to be defined already occurs in the definiens. For a detailed discussion cf. e.g. Balzer and Tuomela (1997).
[120] Bicchieri & Green (1999) acknowledge this character of Lewis', and consequently also e.g. Aumann's, analysis (p. 191, fn. 7). They say that common knowledge and mutual belief can be give analogous definitions. Philosophically, this is a problematic move to make for obvious reasons. However, this trait in Lewis' analysis can be appreciated from the point

(2) **A** indicates to everyone in **P** that everyone in **P** has reason to believe that **A** holds, and
(3) **A** indicates to everyone in **P** that *something **X** is the case*. (Lewis, 1969, p. 56)

Lewis calls the state of affairs, **A**, the basis for common knowledge in **P** that *something **X** is the case*.[121] Lewis (p. 57) states several bases for common knowledge: agreement, salience, the precedents and past conformity, all of which are considered problematic in the literature.[122]
The conditions, (1) - (3), of **(SC)** can, accordingly, be said to provide the basis for common knowledge of social conventions. That being the case, when Lewis' definition of common knowledge is applied to the **(SC)** accordingly, the ***X*** in *something **X** is the case* may be interpreted as forming the content of the left-hand side of the definition. That is, ***X*** = "[a] regularity **R** in the behavior of members of a population **P** when they are agents in a recurrent situation **S** is a convention".[123]

Lewis, thus, defines social convention in part in subjective terms. That is, without reasonable (mutual) beliefs no conventions, and consequently no norms or institutions, would exist. Ontologically speaking, they are, thus, essentially observer relative. However, Lewis requires that, besides that the conditions (1) - (3) are commonly known (or, in this case, mutually believed), the conditions must be true.[124] I take this to mean that the conditions are meant to correspond to reality in a sense that there is a prevailing practice for which conditions (1) – (3) of the (**SC**) hold. The truth and falsity of the conformity to regularities, the expectations vis-à-vis this conformity, and the preferences of the agents are not just matters of opinion or attitude of the observer, and when the conditions do correspond to reality, then the material condition for the existence of a social convention is fulfilled. That is, social conventions, norms and institutions as social facts are anchored to reality instead of being creations of an individual mind, although they could not exist in the absence of mental activity, since preferences, expectations and conformity do not emerge independently of agents.[125]

---

of view that what falls under class of beliefs in individual terms turns out to be knowledge in collective terms. Mutual beliefs about social facts, in a sense, compose those facts. Cf. e.g. Searle (1995) for discussion.
[121] For more detailed discussion, see Lewis (1969), pp. 56-57.
[122] Cf. Gilbert (1990), and Miller (1990).
[123] Cf. Lewis (1969), p. 56.
[124] Of course, when it is known that something X is the case, it must be true. But e.g. the truth of (1) – (3) of the (**SC**) does not yet guarantee it being common knowledge that (1) – (3) of the (**SC**).
[125] Note, however, that social conventions are epistemologically a bit extraordinary. What counts as indivudually subjective transforms on the collective level into objective when preferences and expectations of practices, such as social conventions, are at stake. When it is mutually believed and witnessed that conformity to a prevailing practice has taken place, these interdependent mutual beliefs amount to an attitude that can be considered as common knowledge in the above sense.

5.2.2 Rationality and Structure

The ontological part of the **(SC)** is constituted by the conditions, (1) - (3), of the definition. They form the basis for social conventions. Let us begin from the bottom. As said, condition (3) points to rationality on part of the agents. It does this by referring to preferences and the game theoretical structure of the situation. Thus, it is implicitly assumed that the corresponding agents are maximizers of the expected utility.[126] Condition (3) claims that social conventions are manifestations of coordination equilibria.[127] By conforming to the course of action that implements coordination equilibrium, namely regularity **R**, the agents succeed in avoiding a coordination problem,[128] and, consequently, the problem of indeterminacy. A crucial point of condition (3) is the object that is preferred: conformity to **R**. The conformity to regularity **R** is preferred over the other options, since it implements coordination equilibrium. A deliberated choice under the circumstances of a coordination problem or any other problem of indeterminacy might put the expectations of a satisfactory solution at hazard.[129]

5.2.3 Conformity and Mutual Expectations

Condition (1) and (2) state the commonly known features of conformity and expected conformity to the prevailing pattern. As condition (3) manifests a modicum of rationality, condition (1) and (2) together manifest a sufficient amount of coordination in the past.

Conditions (1) and (2) are best explained by beginning from the point of view of the (**IR**) presented in the previous chapter. The (**IR**) provides a recursive mechanism of inducing expectations about the behavior of agents in part on the basis of past behavior. Going along with the inducements

---

[126] At this point it is not necessary to point out that the agents are assumed to boundedly rational, as well.

[127] David Lewis defines (1969, p. 14) a coordination equilibrium as a combination of strategies in which no one would have been better off had any one agent alone acted otherwise, either himself or someone else. Coordination, in turn, means interdependent choices towards outcomes with coinciding interests. The notion of coordination was, to my knowledge, introduced by Schelling (1960, Chapter 4). In general, people coordinate their actions when they expect each other to do their parts in reaching a certain goal (not necessarily the same or even a common goal) and they expect each other to share the corresponding expectations. Furthermore, the suitable interaction between the agents in pursuing their goals is relevant for the coordination to take place.

[128] That is, by conforming, the agents avoid the problem of making a definite inference about what option to choose. Lewis' (1969) characterization of the coordination problems treats them as situations of interdependent decision-making by two or more agents in which coincidence of interest predominates and in which there are two or more proper coordination equilibria. A proper equilibrium, contrasted with mere equilibrium, in turn, is a solution such that each agent likes the corresponding combination of strategies better than any other combination he could have reached, given the others' choices. A mere equilibrium, on the other hand, is such that each agent likes it at least as well as any other combination he could have reached given the others' choices (Lewis, 1969, p. 22).

provided by this simple recursive mechanism paves the way for choices that under a relatively short sequence may appear as regularity in the corresponding behavior. When the sequence of the game goes beyond the recursive capacities of the agents,[130] the rational agents tend to conform to the prevailing pattern that is inferred from the past events of the recurrent situation rather than at each round making a deliberated choice.

The reasons that speak for this conjecture are evident. To begin, the ordinary agents are assumed to have restricted capacity to make inferences. It is, thus, reasonable to expect that in order to save that capacity for novel and unexpected problems they adopt apparently universally applicable rules of thumb and patterns of behavior, which they tend to follow in familiar recurrent situations. It is plausible to expect such conformity to emerge when the sequence goes recurrently beyond the horizon of the deliberative capacities, since in a recurrent situation it is costly and, in fact, frustrating to make the same recursions time after time. So, when the sequence goes recurrently beyond the horizon, it could be expected that the agents generalize the information gained by the means of the (**IR**) to a rule of thumb or a rule of pattern. In short, conforming to a pattern is a move in a recurring situation such that it is made without present deliberation and it is based in part on precedent events and previous deliberation that may or may not be forgotten by the agents by the time of the present move.

Conformity to the pattern means especially that during present and subsequent rounds the agents act according to the pattern without present deliberation or reconsideration of the choice that has led to the present situation. That is, no presently deliberated choice is made when an agent conforms to a convention. Conformity, in contrast to choice, may be said to represent an idea of *blinkered rationality*. Conforming agents are, so to speak, supplied with blinkers that under given circumstances facilitate social interaction and leave room for deliberation and reconsideration of more important issues. These blinkers may bring about an impression that there never was any choice to be made, and the one and only right thing to do was to conform. Still, the corresponding action may be considered as rational, since according to condition (3) the agents succeed in satisfying their desires by acting accordingly. I take it that the need for acting according to such a generalization is in part determined by the assumption of bounded rationality and restricted capacities.

Condition (1) of Lewis' definition states the conformity to a behavioral pattern. As stated above I take it that this conformity is *de re* based on a generalization of information in part provided by a simple mechanism such as the (**IR**). That is, when it is said that everyone conforms to a pattern it means

[129] Cf. e.g. Gilbert (1990) for discussion. Since the situation presents a problem of indeterminacy, a deliberated choice without taking into account the precedents and, especially, the past conformity reintroduces the problem of indeterminacy that conforming to a convention, norm or institution resolves.

that everyone makes a move that is not presently deliberated, and, possibly, no one has taken under consideration whether that move had any real alternatives. Condition (2) refers to the reciprocal understanding about the prevalence of the pattern and the conformity to that pattern. Especially, the agents expect that a sufficient amount of coordination has already been witnessed in order for them to expect that the regular pattern to conform has formed. The conformity is, thus, based on a generalization of a chain of induced choices and expectations about the corresponding future behavior, and mutual expectations of each other conforming to the corresponding pattern.

5.2.4 Preliminaries for Social Reasons

Lewis' analysis can be said to set forth the basis for the construction of social reasons, in that it refers to interdependent expectations provided by concordant preferences and experiences of mutual precedents. The interdependence of expectations points out the idea of mutual responsiveness of beliefs: the beliefs (or expectations) of the agents are quantified over each other's beliefs, and this is known by each agent.[131] This interdependence is efficacious in the sense that it ties the expectations together in part to form a common basis of reasons for a certain type of move. The common basis of reasons in question is a robust, commonly held basis for action that emerges in part from the precedents and the common knowledge of rationality, and it provides reasons for making the expected move in part because it is commonly held.[132] Thus, in part because of this complex of interdependent expectations and corresponding previous moves, everyone has a commonly held basis for reasons for making the corresponding conforming move. This feature, however, is left implicit in Lewis' analysis. An additional condition is required: (4) (1) in part because of (2) and (3). That means that the conforming move must be made in part by virtue of its being based on the correct reason.[133]

Three features are crucial in creating the system of interdependent expectations. Firstly, the corresponding move must be rationally conceivable. The conceivability principle is implicitly expressed in condition (3) by referring to the preferences of everyone. A move is rationally conceivable if it is not presently dominated by any other move. Since uniform conformity to **R** is coordination equilibrium, or equilibrium in general, it is a rationally conceivable option. The conceivability principle as such does not

---

[130] I take that this is true given the informal presentation of the (**IR**) and that its entailments for the ordinary agents hold.
[131] Cf. the formation of the common knowledge.
[132] Note that making the expected move in part because it is commonly held says more than the reciprocal expectations in a strategic interaction. In the strategic interaction each party tries to be responsive to actions of the others, and vice versa, but this is not necessarily for making an expected move. Cf. the zero-sum games. Bratman (1992) discusses a very similar issue. For a discussion of interdependent decisions, see Schelling (1970), Chapter 4.
[133] Cf. also Tuomela (1995) for a discussion of the proper social norms.

provide any resolution to the indeterminacy problem at hand - it merely points out that there is an available option among several alternatives, and any of the rationally conceivable options will do.

Secondly, a confidence principle must be satisfied in order for interdependent expectations to emerge and persist. According to Lewis (p. 25) an agent has a decisive reason to do his own part if he is *sufficiently confident* in his expectations that the others will do theirs. The confidence principle is implicit in condition (2) of Lewis' analysis: everyone expects everyone else to conform. In order for an agent to rationally hold such expectations, a sufficient confidence must obtain. The prevailing social conventions, at least, satisfy the confidence principle.

Thirdly, besides being sufficiently confident that everyone else will do their own part of the rationally conceivable move, an agent reproduces the prevailing pattern by conforming to it, as well. This conformity must, however, be implemented in part for the correct reason. This feature is expressed by condition (4) introduced above.

Each act of conformity, in turn, adds to a general conformity that paves the way for a path that can be considered self-enforcing if the interdependent nature of the behavior is taken for granted. That is, interdependent expectations make a path salient, and following this path leads to a desirable result from the point of view of the community, as it resolves the problem of deliberation as well.

This argument is consistent with the notion of nested systems of beliefs and expectations presented by Ruben (1985).[134] The existence of a nested system is a *necessary* and *sufficient* condition for social reasons to emerge and endure. According to Ruben, a property is social if its instantiation entails the certain kind of *relation* between individuals. Furthermore, such a social relation exists if and only if (1) there are interdependent expectations about the actions of agents (e.g. conformity), (2) there are interdependent higher-order expectations about beliefs and expectations (e.g. that the agent are expected utility maximizers), (3) there are some descending reason-relations among these beliefs, expectations and actions, so that sometimes what agents do is a consequence of their beliefs or expectations or actions about what other agents expect that they will do, and (4) all this is common knowledge.[135] The reasons for conforming satisfy the requirements of an existing nested system. The third condition of Ruben's definition refers to the confidence principle. The second condition refers to the conceivability principle. The first condition refers to an existing path. All these conditions are subject to the postulate of common knowledge.[136]

An important result of realizing the connection between the accounts of Lewis and Ruben is that

[134] Cf. also Tuomela's (1995) analysis of proper social norms. Both accounts, however, lack the feature that is expressed as condition (4) (1) in part because of (2) and (3).
[135] Cf. Ruben (1985), p. 114.

social conventions are *possibly* not reducible to mere beliefs, in contrast to what the standard game theoretical approach tends to claim. Ruben (pp. 119-127) provides a couple of convincing arguments in support of this result. In addition, assuming that social conventions yield social reasons, I am inclined to accept that social reasons are irreducible to individualistic mental properties, but also that a social reason would not emerge or be sustained without the basis of these mental properties.[137] Especially, the mental properties *in toto* must be *functionally related* with each other in order for the social reasons to hold.[138] Even though it provides the basis for social reasons, Lewis' analysis calls for several clarifications. My main argument is that seeing social habits as equilibria leads the account astray. Besides giving a more adequate picture of social habits, conventions and norms, the view I am proposing resolves also the problems with accounting for the habits under circumstances of conflict. Seeing the issue in terms of funk games clarifies the notion of conformity by referring to the notion of path as a solution concept. Consequently, the scope of the account becomes wider, since e.g. the norms of cooperation are explicable in terms of a self-enforcing path rather than in terms of equilibria. I shall begin the discussion with the issues of scope and applicability of the account, and then I shall defend the notion of the funk game as a clarifying device in accounting for rational behavior in social circumstances.

5.2.5 Restricted Applicability of the Initial Account

In the light of current understanding, Lewis' initial account is too restricting, for it is obvious that social habits should apply to a much wider range of situations than those representing merely solutions to coordination problems.[139]

Andrew Schotter (1981) realizes the restricted applicability of Lewis' account, especially that it

---

[136] The close relationship between Lewis and Ruben is also affirmed e.g. by Gregory Currie (1986).

[137] Cf. Currie's (1986, p. 131) proposition that social properties are supervening on individualistic ones. Using reductive strategies, as e.g. Bicchieri (1990, 1993), Schotter (1981), Ullmann-Margalit (1977) and Lewis (1969) do may still help to clarify the issue, but it does not mean that those who apply reductive analyses are committed to *reductionism*. Cf. Grice (1989), e.g. p. 351, for discussion.

[138] This is a crucial requirement for pointing out the non-reductive character of e.g. social conventions. Following Currie (1986) p. 130, as a car is not merely a set of its mechanical parts, especially when the parts are distributed around the globe, separate individual beliefs and expectations do not amount to a social convention, especially when they are not functionally related with each other. That is, beliefs and expectations need to form a certain structure, in which precedents and future expectations are concordant with the present moves, and all this is common knowledge. Cf. also Bratman (1992) for a similar argument on a different aspect of shared cooperative activity. See even Grice (1989) and his discussion of the Cooperative Principle in Chapter 2.

[139] Ullmann-Margalit (1977) generalized Lewis' view to apply e.g. in the case of PD-like problems. Cf. also Schotter (1981), and Bicchieri (1990). Aumann (1974) introduced the notion of correlated equilibrium, which he claims to be a generalization of Lewis' ideas. Cf. also Vanderschraaf (1995). The evolutionary game theorists apply Lewis' ideas even in much wider range. Cf. e.g. Sugden (1986), Young (1993) and Vromen (1998).

is restricted to the coordination problems. He points out examples, e.g. property rights, that represent social conventions or norms,[140] or to be more precise, social institutions that do not fall into the class that is analyzable in terms of pure coordination games, but in terms of **PD**-type problems. Schotter appears to believe that enlarging the scope of Lewis' analysis succeeds by adding a further condition to Lewis' definition of social conventions. Schotter proceeds as follows:

(**SI**) A regularity **R** in the behavior of members of a population **P** when they are agents in a recurrent situation **S** is an *institution* if and only if it is true that and is common knowledge in **P** that
(1) everyone conforms to **R**;
(2) everyone expects everyone else to conform to **R**; and
either (3) everyone prefers to conform to **R** on the condition that the others do, since **S** is a coordination problem and uniform conformity to **R** is a coordination equilibrium in **S**;
or (4) if anyone ever deviates from **R** it is known that some or all of the others will also deviate and the payoffs associated with the recurrent play of **S** using these deviating strategies are worse for all agents than the payoff associated with **R**. (1981, p. 11)

For the sake of argument, let us take this definition for granted, except for one correction. I think the presentation of condition (4) is not adequate. It should be stated as follows: "*Everyone prefers to conform to **R**, since* if anyone ever deviates from **R** it is known that some or all of the others will also deviate, and the payoffs associated with the recurrent play of **S** using these deviating strategies are thus worse for all agents than the payoff associated with **R**." That is, pattern **R** is preferred over the pattern of deviation, since by deviating the pattern collapses, which is worse for all than when everyone conforms.[141] In fact, this is a formulation that I believe that Schotter meant. Of course, the correction appears to be problematic e.g. from the point of view of explicating institutions that arise in the standard **PD**-type situation, but the following discussion will show that it leads in the right direction. Note also that the (**SI**) requires a similar correction as did the (**SC**), namely (5) (1) in part because of (2) and (3) and/or (4).

[140] Tuomela (1995) e.g. on pp. 16-22 for instance, discusses the common features of conventions and social norms. Cf. also his definition of social norms (pp. 22-28), especially (*SN*).
[141] Nicholas Rowe (1989) has given an alternative formulation. He suggests (p. 22) that social institutions are constituted by agents that follow *rules of action* and believe that others will follow *rules of action*. Furthermore, it is rational to follow a rule of action because by doing so an agent can influence other agents' expectations of his future actions, and can thereby influence their actions to his advantage. However, Rowe's treatment leads to the same problems as the accounts under discussion. That is, he is committed to a notion of rationality that does not leave room for enlarging the scope in the intended sense. This will become evident in the course of the present discussion.

Schotter bases the expansion of scope expressed by condition (4) to Ullmann-Margalit's (1977) taxonomy of social norms. Ullmann-Margalit suggests that social norms can be divided into at least three classes, viz. (i) norms of coordination, (ii) norms of partiality, and (iii) **PD**-norms, and that the norms falling into each class can be given a game theoretical presentation. Preanalytically, norms under classes (i) and (ii) can be presented in terms of (**SC**) as far as they represent the resolution of coordination-type problems, and more generally, the resolution of problems of indeterminacy. Norms under class (iii), on the other hand, are taken into account by condition (4) of the (**SI**). This is the cornerstone of Schotter's contribution to expanding the scope of applicability of the account initiated by Lewis.

To illustrate, condition (3) of the (**SC**) can be presented as the game of pure coordination (**CC**), but also the game of coordination with conflicting interests (**BS**) applies. Condition (4) of the (**SI**) is meant to capture situations that can be presented as the **PD**:

(**CC**):

	C	**D**
C	(1,1)	(0,0)
D	(0,0)	(1,1)

(**BS**):

	C	**D**
C	(2,1)	(0,0)
D	(0,0)	(1,2)

(**PD**):

	C	**D**
C	(3,3)	(1,4)
D	(4,1)	(2,2)

Figures 5.2.1 (**CC**), 5.2.2 (**BS**), 5.2.3 (**PD**)

According to Ullmann-Margalit (1977), **PD** norms are of central importance not because they enlarge the scope of Lewis' account, but because **PD**-structured situations constitute a type of situation, which is prone to generate norms.[142] By adopting such norms the participants in such a situation may guarantee a satisfactory outcome. Furthermore, according to Ullmann-Margalit (pp. 12-13), **PD** norms are norms *par excellence* as far as norms that fall under the following preliminary description are concerned: "A social norm is a prescribed guide for conduct or action which is generally complied with

[142] Ullmann-Margalit, (1977), p. 22. Cf. also the folk theorem, and the discussion of Kandori (1992).

by the members of a society," especially the norms of obligation.[143]

More specifically, the norms under discussion are rules of behavior that are backed by sanctions. The function of sanctions is to transmit information about the expectations of (or attitudes regarding) how one should behave in the corresponding situation. For instance, approval and disapproval of certain acts may serve to signal expectations about correct behavior. As such they do not necessarily restrict or inhibit any actions, nor do they affect the preference structure. Rather, they constitute a guide for correct behavior in complex social situations. That is, sanctions point out the possible norm in a given situation.

5.3 The Equilibrium Account

Ullmann-Margalit (p. 29) proposes that in order to generate a **PD** norm, a stabilizing device that eliminates the temptation to deviate is needed, which, in Ullmann-Margalit's terms, basically means that the mutually satisfactory outcome is "equilibrated" by means of this device. Thus, Ullmann-Margalit proposes an external resolution by bringing about restraints on choices or by changing the order of preferences. Schotter, on the other hand, tries to account for an *organic* way of explicating cooperative behavior under **PD** circumstances. That is, Schotter's aim is to seek a self-enforcing resolution, in contrast to one which is externally authorized. Game theory usually seeks the resolution from this direction.[144] In this view, precedent plays an important role, as looking at the (**SI**) clearly shows[145]: it refers explicitly to the conformity to regularity in the behavior of the members of the group, and this regularity is a matter of evidence about the past behavior. Despite of this character of the initial analysis, some authors are apt to dismiss it in favor of explicating the resolution in terms of (Nash-) equilibria.[146]

The question of the present account is, as was already pointed out, how to account for the self-

[143] For more detailed discussion, cf. Ullmann-Margalit's treatment of Hart's (1961) formulation of the characteristic features of the subclass of *norms of obligation.* Hart (pp. 84-85) in fact speaks about rules of obligation, which Ullmann-Margalit dubs into terminology of norms. The characteristic features are:
(i)[Norms] are conceived and spoken of as imposing obligations when the general demand for conformity is insistent and the social pressure brought to bear upon those who deviate or threaten to deviate is great.
(ii) The [norms] supported by [a] serious pressure are thought important because they are believed to be necessary to the maintenance of social life or some highly prized feature of it.
(iii) It is generally recognized that the conduct required by these [norms] may, while benefiting others, conflict with what the person who owes the duty may wish to do.
As Ullmann-Margalit points out, the main elements of the class of norms under discussion are significant social pressure for conformity to them and against deviation from them. She claims that the **PD** norms respond to all three features of Hart's formulation (Ullmann-Margalit, (1977), pp. 12-13).
[144] Cf. Bicchieri, (1990, 1993), Kandori (1992) and Taylor (1987).
[145] Cf. the discussion of the (**SC**).
[146] Cf. Bicchieri (1990, 1993), Lewis (1969), and Ullmann-Margalit (1977).

enforcing character of the norms under observation. The outlook studied here is condensed by Bicchieri (1990 and 1993, Chapter 6). She claims that social norms are clusters of self-fulfilling expectations that can be broadly defined in terms of equilibria. In order to be self-enforcing or self-fulfilling in the given sense, social norms must be given an explication in terms of equilibria, and to be more exact, in terms of Nash equilibria.[147] In order to account for the self-enforceability of a norm, Bicchieri (1993, 223) claims that a *theory of learning* (i.e. a theory of belief-formation) is needed. Since Bicchieri bases her account of norms of cooperation on the notion of equilibrium, I think that she fails to provide an adequate introduction to a theory of learning. This is due to the property of Nash equilibria that they are reducible to their normal form and thus fly in the face of the notion of learning.[148] Consequently, there is no mechanism that could enforce a norm of cooperation especially under **PD** conditions. Furthermore, a resolution that is expressed in terms of perfect equilibrium refers to ideal rationality rather than bounded rationality. Given ideally rational agents, introducing norms, conventions and customs that facilitate interaction appears to be redundant. Consequently, the norms of cooperation will be left without an adequate explanation. This is a problem of the equilibrium account in general.

There are plenty of possible ways to account for cooperative behavior in circumstances of conflict. What I call the equilibrium account holds that social habits that, among other things, resolve social conflicts should be explained in terms of equilibria. Since, however, perfect equilibrium as a resolution to a conflict situation is generally considered suspect, certain extra requirements are needed. Bicchieri (p. 225), following Kreps et al. (1982), tries to save the equilibrium account for cooperative behavior by *relaxing the postulate of common knowledge of rationality.* Bicchieri feels that this is the only way to overcome the problem presented by the **BI** paradox.[149] In short, since the common knowledge of rationality and making a cooperative move appear to be inconsistent with each other on the grounds of the **BI** argument, one of them must be given up. Since we are accounting for norms of cooperation - that is, rules that enforce cooperative moves - common knowledge of rationality must be sacrificed.[150]

Kandori (1992), on the other hand, studies the conditions of minimal information of making a

[147] Cf. Bicchieri (1993, 222). Since the concept of Nash equilibrium is probably the most frequently applied solution concept, Bicchieri claims that it is important to spell out the conditions of the social norms under which the Nash equilibria may be expected to obtain. Cf. also Bicchieri (1990), p. 840.
[148] Cf. e.g. Binmore (1987), who criticizes the notion of Nash equilibrium for being static and expresses a desire for a more dynamic notion and theory of learning.
[149] Cf. the discussion in the previous chapter. Cf. also Bicchieri (1989) and Pettit & Sugden (1989) for discussion.
[150] According to the standard requirements of game theory this is apparently the case. If the players are supposed to follow the principles of admissibility and uniqueness (interpreted strictly and narrowly), and they expect each other's rationality, then the rational conclusion to make seems to be defection, which naturally is inconsistent with a postulate of first-round cooperation.

justified move under the given circumstances. He claims that in order to sustain cooperation, no more than an agent's personal experience need be known. In order to show that a social norm can be accounted for in terms of equilibrium Kandori must also show that the corresponding agents have an incentive to follow the norm after any history.[151] He constructs a mechanism of dual punishment and assumes that each agent is provided with a *publicly observable label.* In a sense, Kandori's resolution is based on selective incentives, since if a person wishes to expect cooperative moves on the part of the others he must be equipped with a suitable label. It, in turn, means that an agent has successfully cooperated with those who have "clean" labels and punished those who don't. That is, only the "right" persons are treated properly.

The third option for the equilibrium account is to reinterpret social habits more strictly while loosening the requirements of the standard solution concept. Following Lewis' insight Robert Aumann (1987) introduces the notion of correlated equilibrium in order to make certain intuitively appealing outcomes available. Peter Vanderschraaf (1995a) applies Aumann's notion and modifies the question of social habits to be applicable only under circumstances of relatively minor conflict. According to Vanderschraaf, social conventions, norms etc. are seen as the *salient correlated equilibria.* Then, circumstances of conflict, especially **PD**s and **CG**s, are ruled out of the scope of analysis concerning social habits. As the aim of Bicchieri and Kandori is to suggest an account for the norms of cooperation, Vanderschraaf seems to give up the whole idea. If, however, the importance of the norms of cooperation is admitted, the standpoint of salient correlated equilibria presented by Vanderschraaf is unacceptable.

In short, the options of an equilibrium account are at least those of giving up the common knowledge of rationality, of applying external devices or of ruling out certain intuitively acceptable options. All of these options are more or less suspect from the point of view of the present approach. Since we already have an alternative account that succeeds in overcoming the problems of the equilibrium account, while being able to give a more adequate explication of social habits, the equilibrium account can be rejected.

5.3.1 Giving up the Common Knowledge of Rationality

Let us first observe Bicchieri's argument for extending the scope of Lewis' account - that is, the justification for **PD** norms, or norms of cooperation that are broadly defined in terms of equilibria.

[151] Cf. Kandori (1992), p. 64.

Bicchieri (1993, p. 221) suggests that social norms can be defined either as cooperative equilibria that are explained in part in terms of repeated interaction and small uncertainty about the rationality of the opponent, or as established behavioral regularity that has emerged through repeated interaction. Either way, Bicchieri (1993, 232) sees social norms as equilibria in the game theoretic sense of being combinations of strategies, such that each individual strategy is a best reply to the other's strategies. This property of social norms is, according to Bicchieri (1993, 232), verified by conditions (2) and (3) of the (**SC**). Bicchieri explicates the notion of norm as follows. Let *R* be a behavioral regularity in population *P*. Then, more generally, *R* is a *social norm* iff *R* depends on the beliefs and preferences of the members of *P* in the following way: (1) almost every member of *P* prefers to conform to *R* on the condition (and only on the condition) that almost every one conforms, too, and (2) almost every member of *P* believes that almost every other member of *P* conforms to *R*. This characterization suggests that social norms are essentially describable in terms of social conventions in the sense of Lewis.

Bicchieri's problem, however, seems to be how to justify the norms of cooperation, if they are at the same time considered as equilibria and mutually acceptable cooperative resolutions to the social dilemma in which a perfect equilibrium solution yields an unsatisfactory result. Consequently, condition (4) of the (**SI**) calls for justification as well. That is, norms of cooperation should satisfy the equilibrium requirement while resolving the problem of unsatisfactoriness. Then, conformity to the norm should directly maximize the expected utility even in the case of an iterated *n*-person **PD**. This result seems to be self-defeating, since aiming at an equilibrium solution is what the direct maximization is all about. When ideal rationality is at stake, as is the case with Bicchieri's analysis, a well-defined unique and admissible solution can be found in finite iterated **PD**s.[152] How, then, can the norms of cooperation be explicated if it is assumed that the resolution is presented in terms of equilibria?

Bicchieri claims, following Kreps et al. (1982), that by relaxing the postulate of common knowledge of rationality, the cooperative patterns become conceivable even in the circumstances of the **PD**. According to Kreps et al. cooperation can result when the players have slight doubts about each other's rationality. An advantage of this approach is, according to Bicchieri (p. 221), that complex reasoning chains of the "I know that you know that I know..." can now be eliminated. The drawback of the approach, according to Bicchieri, is that it requires the use of *ad hoc* hypotheses, such as that players have common knowledge of all the possible types of players they might face and of the relative frequency with which each type may appear. Especially, it must be assumed that the probability that one

applies tit-for-tat is given and it is assumed to be common knowledge among the players.[153] According to Bicchieri such an assumption must be made, since assuming that the players are capable of making an *accurate* probability assessment about each other's type is at odds with what happens in real-life interactions.[154] But Bicchieri's proposal would mean that cooperation takes place when the players have an interest in acting *as if* they were being cheated about each other's rationality, when they in fact are not. Furthermore by adopting this "as if" strategy the players hope to induce a further cooperative act by the opponent. Then, “rational” cooperation in **PD** situations may occur, and a “rational” basis for the emergence of norms of cooperation is laid down. To my mind norms of cooperation are, then, based on self-deception, since, then, the agents base their moves on expectations which they know to be incorrect.

Now, I do not think that giving up the common knowledge of rationality is necessary for lightening the machinery as far as complex reasoning chains are concerned.[155] Nor do I think that one needs to use the above kind *ad hoc* hypotheses in order to fill the gap created by giving up the common knowledge postulate. It is important to understand the inherently incomplete nature of rationality in order to appreciate the view I am proposing. I do not think that rationality increases when the recursions of beliefs are stretched further and further. It might mean fewer mistakes, though. On the other hand, assuming bounded rationality and restricted capacities does not rule out the postulate of rational agents being maximizers of the expected utility. Instead, slight doubts about the common knowledge of rationality would mean to cast doubt on this feature of the agents, which would also mean difficulties in accounting for norms of cooperation in terms of equilibria. The assumption of bounded rationality guarantees machinery light enough to avoid monstrous reasoning chains while appreciating rationality in the presence of mistakes, as well. Furthermore, no *ad hoc* hypotheses are required, since when the inherent incompleteness of rational agents is appreciated, there is no need to expect that the probability assessment of each others type is always correct. In addition, when the conformity to a social norm or convention is at stake the agent can expect certain behavior on part of the other agents, and these expectations can be expected to be based on an approximately correct

---

[152] Note that the equilibrium account defended by Bicchieri requires the notion of ideal rationality. This requirement becomes evident especially when the group size is increased and the agents need to take more information into account.
[153] Cf. Bicchieri (1993), p. 226.
[154] I take this to mean that the players suffer from the lack of capacity to make the correct probability assessment, which makes such an assumption dubious.
[155] Cf. also alternative approach provided by Balzer and Tuomela, (1997). The complex reasoning chains are condensed into a fixed-point argument in which the defined property of the common knowledge postulate is included in the definiens. The fixed-point argument may facilitate the treatment of collective beliefs, but the inferential character of holding a rational belief seems to be missed. At this point it is not possible to offer any argument for this insight. It

probability assessment in the given circumstances. This assumption can be said to be verified e.g. by the common knowledge of conditions (1) – (3) of the (**SC**) supplemented with (4) (1) in part because of (2) and (3).

Although Bicchieri's course of reasoning leads the account astray, I think the basis of her approach is solid in trying to reduce the Bayesian *omniscient* tones of the approach. For instance, she proposes (1990, 846) that cooperation becomes less surprising if we think that rationality, far from being a specific, clear-cut mode of action, is an inference of the best choice, given the beliefs we have about the circumstances of the game. This line implies the seeds of robustness and resilience that I think are essential features of rationality. Given that we are dealing with boundedly rational agents with restricted capacities of processing beliefs, it is not necessary to introduce an aspect of doubt to each other's rationality in order to make cooperative moves eligible. Thus, the expectations of rationality do not have to collapse when "unexpected" behavior is witnessed, and when the information about the circumstances is updated, the agents are capable of revising their course of action.

However, a big leap is yet to be made from an inference of the best choice to conformity to a prevailing pattern. That is an issue that Bicchieri totally dismisses. Even if the organic account that aims at enforcing an endogenous resolution in terms of equilibria does not seem to provide the desirable result, the external correlating device may have the desired effect. The external correlating device may come in several alternative forms. It may be an arbitrary artefact, such as a coin toss before playing the game, or a label on the player's forehead, or experiences of common past events. As an example of the approach that goes in terms of arbitrary and artificial external correlating devices, let us study Kandori's (1992) account of labels.

5.3.2 Applying the Selective Incentives

Michihiro Kandori (1992) claims that for mutually beneficial cooperation to emerge and be sustained, systematically revised labels[156] suffice to provide information needed to make a justified move. This is very close to the idea of the *ad hoc* hypotheses defended by Bicchieri. Only now, only personal experiences and checking the label before the play are what counts. In a sense, the agents apply selective incentives in order to sustain the cooperative pattern in the community.[157] Besides providing selective incentives by means of labels, Kandori (p. 64) assumes that sustainable social norms must

---

suffices to say here that the essential incompleteness of rational inferences resolves the problem of infinite regression as well as the problem of *ad hoc* postulates.

[156] E.g. reputation, membership, license, etc.

provide proper incentives to the agents in every respect, and that the labels provide a correlation device for equilibrating the outcome in the intended sense.

The incentive problems of the agents are treated with a dual punishment mechanism.[158] That is, the sanctions that support the norm work two ways: they promise a mutually beneficial outcome for conformers - after all, the cooperative *pattern* that the norm recommends is preferred over a non-cooperative *pattern*[159] - and they threaten defectors with retaliation. Furthermore, Kandori provides a further incentive for conforming: those who fail to punish are themselves punished in turn. All this information and only this information is recorded in the labels. In this way, Kandori claims, social norms represent a robust equilibrium that sustains any mutually beneficial outcome, such as universal cooperation in the **PD**.

However, it is evident that even if Kandori's analysis captures condition (4) of the (**SI**) and succeeds in specifying it, and consequently succeeds in explicating how norms of cooperation were enforced, it does not succeed in showing that an arbitrary norm of cooperation is an equilibrium outcome. Recall that condition (4) says that everyone prefers to conform to regularity **R**, since if anyone ever deviates from **R** it is known that some or all of the others will also deviate, and the payoffs associated with the recurrent play of **S** using these deviating strategies are worse for all agents than the payoff associated with **R**. A drawback of using an external device, such as labels, is, then, that it is not obvious whether the cited social norm succeeds in realizing an equilibrium outcome.[160] This means difficulties in warranting the equilibrium account. Especially, the **PD** norms still seem to sustain an out-of-equilibrium mode of behavior, rather than an equilibrium outcome.

If we are to accept Schotter's (1981) initial argument, which is supported also by Bicchieri (1993), that an ultimate resolution must be sought from organically or spontaneously emerging options, an external authorization of e.g. a dual punishment is ruled out. This kind of external device may be applicable in higher-order institutional constructions. However, as my study is seeking to justify such higher-order constructions, and to my opinion the justification must begin within lower-order terms, external authorization à la Kandori or Ullmann-Margalit is not available. Anyway, as the emergence of norms is a matter of spontaneous order - a form of coordination that does not presuppose previous agreement[161] - it suffices to show whether the emergence of the norm is rationally conceivable. The rest

---

[157] Cf. also Olson (1968) for a discussion of solving social dilemmas by means of selective incentives.
[158] Kandori (1992), p. 74. If a non-conformer meets another non-conformer, they mutually minimax each other. If a non-conformer meets a conformer, the conformer minimaxes the nonconformer, but the non-conformer is not supposed to minimax the conformer. According to Kandori, he is supposed to "repent".
[159] This is supported e.g. by Bicchieri (1993), p. 226.
[160] Cf. Kandori (1992), p. 65.
[161] Cf. Bicchieri (1990), p. 861.

is done by natural selection in the course of the evolution of social habits. In the absence of the external device that enforces the norm, the incentive problem still causes difficulties in sustaining the cooperative pattern. Taking condition (4) of the (**SI**) for granted, a norm of cooperation under the given definition is far too unstable to constitute an equilibrium. Consequently, in order to save the equilibrium account, the notion of equilibrium requires further specification.

In order to hold that social norms, in fact, can be presented in terms of equilibria, Kandori is compelled to employ an essentially weaker equilibrium concept than that of Nash equilibrium. According to Kandori (1992, 71), then, a combination of desirable properties is required. Roughly put, the properties can be stated as follows:

(i) The necessary information for making decisions is transmitted by labels;

(ii) an agent need not know the fine details of the information structure at hand in order to make decisions that are in accordance with the norm;

(iii) an agent need not know the number of participants nor their relations to each other in order to verify whether the norm is an equilibrium;

(iv) small mistakes are ignored in a way that guarantees the stability of the norm, and after *punishing for* a big mistake the equilibrium path is resumed; and

(v) for a social norm to be effective, the rule must be simple enough. A test of simplicity could be for instance the number of available options that are in line with the norm.[162]

In short, social norms are robust equilibria that are in part based on selective incentives (i.e. such labels as credit cards, club membership etc.) and resilience of the equilibrium path.

Although selective incentives as such do not belong to the means of spontaneous order and so are not of interest in this essay, Kandori's analysis has some interesting features. One central feature is the use of an external correlating device. It does not always mean external authorization of the present move. Although social norms may be enforced by means of an exogenous correlating device, this enforcement may still be spontaneous and organic rather than a product of design. For instance, correlating expectations and moves by means of external *events* may work as a spontaneous correlating device. External events are understood to consist of precedents - previous choices, emerged regularities etc. Then, however, the notion of deliberated choice loses some of its force. The moves of the agents seem to be determined by the information extracted from the labels rather than from consideration of all the relevant options. But under these circumstances we are no longer dealing with agents that are equipped with ideal skills and perfect capacity. At least then the introduction of labels would appear to

[162] Cf. the discussion in Kandori (1992) pp. 71-73.

be redundant. Consequently, the notion of equilibrium turns out to be problematic as a solution concept, and conformity to a known regularity gives a better description it.

5.3.3 Weakening the Solution Concept

A good question is why social habits should be explicated in terms of Nash equilibria, when in an important class of games Nash equilibria yield unsatisfactory results, even if the results are unique? Furthermore, why should social habits be explicated in terms of Nash equilibria, when in an important class of games the social habits appear to point out an out-of (-Nash)-equilibrium mode of behavior in order to guarantee mutually satisfactory outcomes? Since the equilibrium account does not seem to deliver the goods by applying the standard solution concept of Nash equilibrium, the view that social habits are broadly defined as equilibria has created pressure to weaken the notion of equilibrium or to define it in more general terms. Robert Aumann (1987) does this by directly referring to Lewis (1969) and setting forth the notion of correlated equilibrium.

The general idea is that, in contrast to Nash equilibria, the agents do not have to know what strategies are actually used by the other agents. Instead, it suffices to assume, as in the case of Nash equilibria, that it is common knowledge that the agents are (Bayesian) expected utility maximizers. That is, the agents are rational in the sense that each *conforms* to the idea that each agent has a subjective probability distribution over the set of all states of the world, and that this is common knowledge.[163] Furthermore, in correlated equilibria, there is no need to randomize the strategies, as is the case with applying the Nash equilibria. Instead, the players choose a pure strategy, which they may base on observation on the same random variable, such as the witnessed coordination in former rounds, or on some other external event.[164] But does the correlated equilibria account rescue the equilibrium approach?

5.3.4 Salient Correlated Equilibria

An indisputable characteristic of correlated equilibria is that they enlarge the scope of Lewis' initial analysis, since the notion of coordination equilibrium is a special case of Nash equilibrium, whereas the

[163] Cf. Aumann (1987), p. 2: He follows Savage's (1954) and Lewis' (1969) theories in defining the notion of correlated equilibrium.
[164] Aumann considers an example of tossing a fair coin, which can be used to determine a symmetric effective outcome.

notion of correlated equilibrium makes plenty of alternative options available.[165] An interesting question, however, is whether this enlargement of scope embarks on accounting for the norms of cooperation in organic terms. A preliminary answer is no, since the only applicable choice in the **PD**-type problems which the norms of cooperation reflect is the unique solution that is also in accordance with the Nash equilibrium. However, when the situation is observed from the point of view of the emergence and sustaining of social habits in a dynamic framework, answering that question becomes complicated. Indeed, by using correlated equilibria, the scope of Lewis' initial account is remarkably enlarged, but the question is whether the enlargement provided by the notion of correlated equilibrium is effective in accounting for the most interesting classes of social habits.

Peter Vanderschraaf (1995a) attempts to account for social conventions in terms of correlated equilibria supplemented with a requirement that aims at satisfying the criterion of public observability. More specifically, Vanderschraaf claims that the notion of salience can be formalized in terms of the correlated equilibrium that satisfies a *public intentions criterion*, (**PIC**): Each agent desires his intended choice of a strategy to be common knowledge. Thus, Vanderschraaf suggests that social conventions (and, consequently, social norms, institutions and other corresponding habits) are explicable in terms of salient correlated equilibria.

Vanderschraaf claims (p. 71) that (i) his account embraces the idea that conventions operate under circumstances of the games of *mutual interests*, in contrast to games of *conflicting interests*. According to Vanderschraaf, Lewis' account fails to do so adequately, since Lewis' criterion of mutual expectations fails to distinguish between situations with conflicting interests and situations with mutual interests. Furthermore, Vanderschraaf claims (p. 71) that (ii) in part because of (i) in his account succeeds in sharpening the notion of salience by means of a (**PIC**), whereas Lewis succeeds in only vaguely referring to the precedents. Finally, Vanderschraaf claims that (iii) his account enlarges the scope of applicable models for conventions, since the class of correlated equilibria that satisfy (**PIC**) is larger than the class of coordination equilibria. Thus, there are games of mutual interest that can be said to represent the circumstances in which social conventions emerge and are sustained, so that they have a set of conceivable outcomes that do not satisfy Nash equilibria but that still satisfy (**PIC**).[166] Those situations are best explained in terms of salient correlated equilibria. All this, Vanderschraaf claims, is perfectly consistent with Lewis' viewpoint, while correcting the major difficulties in the initial analysis.

The claim (i) is targeted to capture the idea presented by Schelling (1960, (reprint in 1970)) and, more recently, by Bratman (1992). It is worth noting that reciprocal expectations as such do not

[165] Cf. the discussion in Aumann (1987) and Vanderschraaf (1995a, 1995b).

adequately account for the emergence of social conventions, since in full-blown strategic interaction rational agents forecast each other's moves without necessarily aiming at a mutually beneficial outcome. Then, especially, Lewis' claim (p. 25), that an agent has a decisive reason to do his own part of the possible coordination equilibrium if he is sufficiently confident in his expectations that the others will do theirs is at stake. Vanderschraaf (p. 69) baptizes the claim the *mutual expectations criterion* (**MEC**) which is presented as follows: "Each agent has a decisive reason to *conform* to his part of the *convention* given that he expects the other agents to *conform* to their parts." According to Vanderschraaf, the problem with the (**MEC**) is that it can be satisfied at any strict Nash equilibrium, and because a remarkable amount of Nash equilibria do not satisfy the requirement of being mutually beneficial, the (**MEC**) is not an adequate criterion for conventions.

Lewis does not, strictly speaking, give any criterion of mutual expectations that can be interpreted merely in terms of strategic deliberation, as Vanderschraaf's critique suggests. Instead, he (p. 25) speaks of each agent doing his part to fulfil a mutual goal (by referring to the notion of coordination equilibrium which, roughly put, means outcomes with coinciding interests) and each agent's expectations that everyone else will do their part. That can hardly be said to satisfy *any* strict Nash equilibrium. Thus, when Lewis' analysis is read benevolently and the basic idea of reaching a resolution to the problem of indeterminacy is taken into account, the norms of cooperation should belong among the facilities for providing resolutions to social dilemmas that arise from conflicting interests in recurring situations. Even Vanderschraaf's own formulation of the (**MEC**) does not support the claim that it can be satisfied at any strict Nash equilibrium, even those that are in conflict with outcomes of mutual interest, since it refers to conformity to conventions and conditional expectations about this conformity. It is clearly the outcomes of coinciding interests to which the criterion of mutual expectations is constrained to apply, and under recurring circumstances, outcomes of coinciding interests that conflict with e.g. Nash equilibria become rationally conceivable, as is well established in the literature since Axelrod (1980) and others.

This means that there is no *a priori* need to exclude certain models of conflicting interests, e.g. **PD**s and **CG**s, as Vanderschraaf is inclined to do, although, in general, conventions provide resolutions to social dilemmas in terms of mutual acceptability and rational conceivability. Even then, the principal limit of applicability is between zero-sum games and mixed-motive games, since the possibility of cooperation, and of cooperative social habits, is enormously increased under dynamic circumstances.[167] Ruling out the models of conflicting interests at the stage game level seems more like an *ad hoc* strategy

[166] Cf. Vanderschraaf (1995a), p. 74.

that attempts to force reality to fit into the model. If Lewis' claim (p. 25) about mutual expectations is read benevolently, the main point is the problem of indeterminacy and its resolution, rather than a matter of the coordination equilibrium and the coordination problem. So, any problem of indeterminable choice that can be seen as a question of striving for a mutually beneficial outcome falls within the scope of Lewis' analysis, even though a literal reading of Lewis' analysis of the preference structure of the situation leads to the abandonment of a crucial class of social habits, especially the norms of cooperation.

When it is granted that ruling out the mixed motive games with conflicting interests, especially iterated **PD**s and **CG**s, is not warranted, the introduction of the (**PIC**) becomes dubious. Suppose for the sake of an argument that the (**PIC**) is satisfied. That is, everyone wishes his choice of strategy to be common knowledge. Then, on the condition that resolving social dilemmas is among the functions of social habits, the norms of cooperation would be unexplicated under the given description of correlated equilibria. Under the circumstances of the **PD** or the **CG** no one wishes his choice of (a mutually beneficial) strategy to be common knowledge under the given description of the correlated equilibria. But this conclusion is inconsistent with the idea that norms of cooperation should be included into the machinery. So, the (**PIC**) must go, and the notion of salience is still left without a proper treatment. Consequently, the enlargement of scope due to salient correlated equilibria is only cosmetic for the present purposes. Norms of cooperation are ruled out, and if they are taken in, Vanderschraaf's notion of salience becomes dubious.

5.3.5 An Equilibrium Account: Concluding Remarks

As the above discussion shows, the equilibrium account does not succeed in delivering the goods that its proponents promise. The equilibrium account embraces a theory that is at odds with the real world in assuming that the agents should preview all possible lines of reasoning the future may bring. That is, the equilibrium account is committed to Bayesianism, according to which rational agents are never surprised. By this I mean especially that the equilibrium account embraces the idea of ideal rationality in terms of ideal skills and the perfect capacity of making the necessary inferences. However, the Bayesian epistemology is far too demanding for an ordinary boundedly rational agent in the complex world we live in. The (**IR**) shows how even relatively low-level deliberative processes can be complex, and their complexity increases when the recursions of beliefs are stretched further and deeper. That is, from the

[167] Cf. Schelling (1970), Chapter 4, for a preliminary discussion.

point of view of ordinary agents the world is hopelessly large in Savage's sense. A player cannot consider exhaustively every contingency that one finds in one's own deliberations and what one expects to find in others' deliberations.[168]

Furthermore, it is clear that because of its Bayesian tones,[169] the correlated equilibrium account does not directly contribute to an account that operates with the notion of bounded rationality and restricted capacities. It is too much to require that ordinary agents would ever deliberate a probability distribution over all states of world, even if this meant only the states that are relevant to the present move. The discussion of the backward induction paradox gives a good indication of the unintuitivity of the Bayesian conception of rationality: a Bayesian agent is never surprised, but ordinary rational agents may often be, without there being any justification for accusing them of irrationality. Despite its Bayesian tones, Aumann's notion of correlated equilibrium is an important contribution to the discussion of social habits. It aims at accounting for action based in part on reciprocal expectations inferred from external events, while requiring that the observations of the agents need not be identical or independent on each other. Especially, Aumann's theory anticipates the signal account in that it assumes that an agent does not have to know the whole system, but has to look for the signals emitted by the other players that add to the posteriors of the agent. Furthermore, Aumann's theory anticipates the idea of a path instead of an equilibrium as a solution concept in that it applies the posteriors about other players' choices, and these posteriors add to the substantive information on which agents (in part) base their present and future choices.[170]

Nevertheless, the equilibrium account does not seem able to give any solid reasons for acting cooperatively e.g. under circumstances of the **PD**, unless we accept that cooperative action is in general based on self-deception. Especially, when norms of cooperation are at stake the equilibrium account faces difficulties on grounds of recommending options that are better explained in terms of irrationality.

The understanding of rational conceivability of e.g. the cooperative pattern in **PD**-type situations turns out to be problematic if the discussion is anchored in the notion of equilibrium, even if it were the notion of salient correlated equilibrium, since it tends to illustrate the circumstances in an unnecessarily static fashion. Understood as an account that operates with an idea of self-enforcing path, in contrast to a self-enforcing equilibrium, the signal account seems to provide a solid basis for explicating social habits.

---

[168] Cf. Binmore (1993), p.330.

[169] I am using the term 'Bayesianism' in the sense of Binmore (1993): I am not denying the use of Bayes' rule. Instead I want to stress that e.g. in the case of boundedly rational agents and restricted capacities it is too much to require that the agents "look before they leap" when the sequence on which they are deliberating goes beyond their recursive capacities. Cf. Savage (1972) for a distinction between a small and a large world.

5.4 An Argument from the Self-Enforcing Path

Many of the current problems would have been avoided if Lewis had formulated condition (3) of the (**SC**) in another fashion - without referring to the notion of *coordination equilibrium.* It is evident that the notion of equilibrium as a solution under the cited circumstances does not correctly illustrate the mechanism behind the conformity to conventions. If we do this, the resolution is ultimately reduced to a *static* form of equilibrium under circumstances in which conformity to a pattern gives a better description of the situation. Conformity to a *social habit* is better stated in terms of a path that is in part "enforced" by past events and that is in part reproduced by mutually expecting a corresponding reciprocal conformity to the pattern that is created by past and present behavior. This means that conformity should be expressed in *dynamic* terms of a self-enforcing path, rather than in static terms of a perfect equilibrium, if it is to be given a correct interpretation.

This feature of conformity does not mean that the normal form presentation could not be interpreted as an extract of the precedent history including all the aspects of learning and expectations of rationality that are typical of conformity in general. However, even if the normal form presentation is an extract of all things that must be considered in a given interaction situation, it does not succeed in adequately presenting a social habit, convention or norm. This point can be clarified with the example illustrated in Figure 5.4.1. In this example, let us suppose, the precedents have led to a situation in which, all things considered, an agent may expect other agent(s) to go either **C** or **D**, and the agent finds himself in a problem of indeterminacy.

	C	**D**
C	(1,1)	(0,0)
D	(0,0)	(1,1)

Figure 5.4.1 A normal form of the game of coordination

This being all the information an agent possesses in the given situation, no corresponding social habit or convention can be said to have emerged, since the situation represents a problem of indeterminacy that appears as yet unresolved. This illustration does not show any sign of regularity in behavior or of being an existing and recognized pattern in force. Accordingly, the normal form presentation is not an

[170] Cf. discussion in Aumann (1987), pp. 8-9.

adequate way of illustrating social habits. Moreover, by adopting the dynamic path view instead of the static, slice-wise, normal form view, the notion of salience gets a natural formalization, as well. Consequently, the main criticism of the normal form view is avoided.[171]

The self-enforcing path is brought about by an existing social habit, viz. a custom, a convention, a norm, or an institution, that can be described in terms of common knowledge and existence of conformity, suitable reciprocal expectations and a rationalizable basis for the conformity and the expectations. A social habit can be defined by slightly adapting Lewis' definition of the social convention as follows:

(SH) A regularity **R** in the behavior of members of a population **P** when they are agents in a recurrent situation **S** is a *social habit* if and only if it is true and is common knowledge in **P** that
(1) virtually everyone prefers to conform to **R** on the condition that the others do, since **S** represents a problem of indeterminacy, and uniform conformity to **R** provides a rationalizable path out of the problem represented by **S**;
(2) in part because of (1) virtually everyone *expects* everyone else to conform to **R**; and
(3) in part because of (2) virtually everyone *conforms* to **R**;

It is important to observe that these social habits take place by *conforming* to certain existing and recognized regularities, and by mutually *expecting* that the others are conforming to them as well. These two features are crucial also in Lewis' (1969) initial definition, but they have often been dismissed in favor of the discussion biased by emphasizing the alleged preference structure of the situation. Especially, *conformity* to a pattern has been often identified with a *choice* in a strategic interaction situation. Distinguishing between these two moves, viz. conformity and choice, is crucial, since conformity refers to a move that is not based on present deliberation, whereas a choice is necessarily a matter of deliberated decision, no matter how banal it may be. I take it that the rationality of conformity is a matter of precedent decisions and the reasons behind those decisions need not be actively contemplated in making the present conforming move.[172] Conformity is essentially based on learning and adopting a routine manner of behaving. Acting according to the adopted pattern facilitates social interactions remarkably and, indeed, resolves certain social dilemmas, as well. Furthermore, a lot of the

[171] Cf. e.g. Gilbert (1990), especially p. 10, where she emphasizes the structure of the situation in favor of an existing convention.
[172] Cf. certain beliefs and desires that may, in a sense, constitute certain principles and facts that are not presently deliberated, although they may have influence on the present moves, e.g. a belief that 134 + 67 = 201, or a desire for making a good living.

capacity of deliberation is kept in reserve for real need when a conforming deed is done.

Condition (1) of the (**SH**) emphasizes the properties of conforming to a pattern by referring to a *path*, instead of leading the account astray by referring to the notion of equilibrium mainly for reasons of accounting for conformity in rationalizable terms. The notion of equilibrium is directly connected to the notion of strategic choice, which, as already noted, does not adequately describe behavior under the description of conforming to social habits, e.g. following a rule or obeying a norm.

The path view has certain welcome properties. Firstly, it is not committed to the restricting constraints that the equilibrium account must satisfy. Especially, the stability criterion entertained by Lewis is unnecessarily strict. The analysis of social conventions that goes in terms of the coordination equilibria, and the Nash equilibria in general, rules out important classes of the study, e.g. norms of cooperation, as the discussion presented above proposes. Furthermore, the equilibrium account is, in fact, tied to ideally rational beings that do not need any customizing of rationality in order to facilitate recurring social interaction. Especially, ideal agents are not disturbed by the unintuitivity of the ongoing unsatisfactoriness of certain outcomes. The path view, instead, succeeds in providing adequately stable outcomes that resolve the problems of indeterminacy, even under circumstances that call for norms of cooperation. In a large world, for instance in a recurrent situation in which the horizon of beliefs goes beyond recursive capacity, when the prospects are promising, a cooperative pattern may be rationally conceivable.[173] But such a large world cannot be captured by means of equilibria alone. Instead, a norm of cooperation may be adequately described as a path that (i) points out a resolution of the problem of unsatisfactoriness, and (ii) resolves the problem of indeterminacy introduced by the resolution of the problem of unsatisfactoriness.[174] The path provides a resolution to the social dilemma by means of a salient option, namely conformity to the prevailing pattern. The stability of the resolution is guaranteed by the existing and recognized pattern itself. Of course, I am not speaking about an absolute stability, but stability related to the current situation. The path reinforces itself as long as the *degree of assurance* (cf. the next section) is such that the prospects are promising enough to guarantee the rational conceivability of the cooperative pattern. Thus, the outcome is stable as far as certain requirements are met. However, the requirements of stability need not be so strict as to satisfy perfect equilibrium.

Before the analysis may proceed, an elaborate study of condition (1) of the (**SH**) seems to be required. The crucial question is how the path view can be accounted for. In Chapter 4, I presented the mechanism of the (**IR**) that at this point turns out to be of essential importance. To summarize, the (**IR**)

[173] I.e. when the degree of assurance allows one to expect cooperative behavior within a reasonable time.

shows how ordinary agents may induce rational expectations of agents' behavior to each other under certain circumstances. Furthermore, the (**IR**) suggests that even in the simple cases the deliberation over eligible options becomes increasingly complex the further the recursions are stretched. Thus, when ordinary agents are involved, it appears to be rational in recurring situations to adopt certain habits that make routine moves possible, instead of deliberating the best move over and over again at each. As agents' skills and capacity are limited, social habits may offer a short cut in managing in such situations.

5.4.1 Accounting for a Cooperative Path

The argument for the self-enforcing path can be approached from the point of view of the (**IR**). The (**IR**) suggests that a *cooperative* move can be eligible even in a relatively short sequence. In short, the (**IR**) states that in a situation in which an agent is not at the position to make a backward induction, the agent may try to avoid the problem of unsatisfactoriness and induce expectations of cooperation to the others by making a cooperative move. Recall that rationality of this act depends on three conditions:
(i) the cooperative act must provide a credible signal for further cooperation on part of the agent;
(ii) the cooperative act must in part be based on a credible signal for further cooperation on part of the other agent(s); and
(iii) (i) and (ii) are commonly known.

Condition (i) says that the cooperative act of the present round must be a signal that cooperation during the next corresponding round from the point of view of the agent is eligible option, and it is an eligible if it promises better prospects than refraining from cooperation at that point. This condition suggests that the *degree of assurance* is an important factor in making a cooperative move. The degree of assurance works pretty much in the same way as the *discount parameter*.[175] That is, a lot value is placed on the sequence close to the present move. This means, especially, that the information that is beyond the deliberative capacities in the light of the present move has only a limited effect on that move. For instance, knowledge of the finite nature of the game and its principal structure does not justify making the deductive **BI** argument. ”Pulling down” irrespective of the expectations of the others' behavior is not warranted, since the prospects, given the manageable information, may promise gains that cannot be

[174] That is, the norms of cooperation are seen as social habits of dual structure in contrast to social conventions, as presented by Lewis. The common denominator for these social habits, be they norms of cooperation or social conventions, is that they solve the problems of indeterminacy.
[175] Cf. Axelrod (1981).

achieved by *maximining.*[176] So, for boundedly rational agents it can be said to hold that the sooner the expected utility of the cooperative move exceeds the expected utility of the present maximin move, the higher is the degree of assurance, and the higher the degree of assurance is, the more *confident* the agents may be in their expectations that others will *conform* to the cooperative path induced by the reciprocal conforming moves. If the agents can overcome the problem of assurance, then the cooperative path would become accessible. The problem of assurance is usually illustrated as a game of assurance, the (**AG**):

	C	**D**
C	(4, 4)	(1, 2)
D	(2, 1)	(3, 3)

Figure 5.4.2 the game of assurance

In this example it is important for agents to be able to overcome their suspicions about each other's willingness to cooperate, their capacities of making the necessary recursions on the nature of the situation and each other's rationality. Otherwise, the strategy of maximin seems to be the best option to the players. It is claimed that rational agents are typically able to resolve the problem of (**AG**).[177] For instance, a game in Figure 5.4.3 is easily resolved.

It is important to notice that the agents may face an **AG**-type problem in certain instances of the Centipede as well as in iterated **PD**'s. The **AG**-type problem depends on the fine details of the structure and the length of sequence of the given situation.

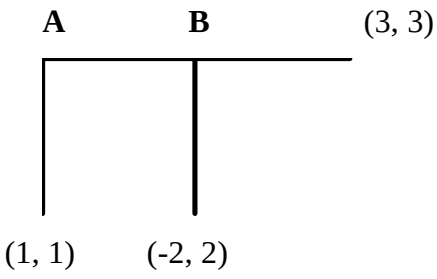


Fig. 5.4.3 the problem of assurance from A's viewpoint

[176] That is, by maximizing the minimum gain.

The game of assurance, the (**AG**) presented in Figure 5.4.2, and the problem of assurance, presented in Figure 5.4.3, can be applied to the game of Centipede under circumstances of satisfactory degree of assurance. Besides the capacity of beliefs, the degree of assurance may be dependent on the agents' cool-headedness, as well.

The degree of assurance is zero when the corresponding move is strictly dominated by an alternative move or when within the depth of everyday recursions a unique and admissible solution is found. That is, the degree of assurance is zero when the move does not satisfy the principle of *rational conceivability*. The single-shot Prisoner's Dilemma and the basic one-unit sequence Centipede (Figure 5.4.4) satisfy these requirements unambiguously, and thereby in those kinds of situations cooperative moves cannot be rationally induced.[178]

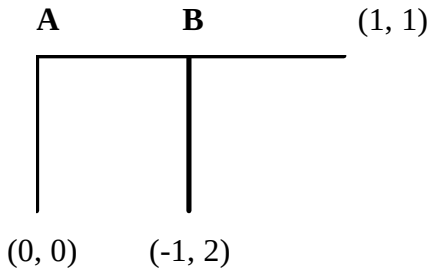


Figure 5.4.4 the one-unit sequence Centipede

The degree of assurance of the cooperative move in Figure 5.4.4 is zero, since agent **A** cannot expect to gain more than 0, which is reached by pulling down rather than moving forward - in which case agent **B**, if rational, will definitely pull down, and **A** would get only -1. There are no prospects for the emergence of cooperation, since the sequence ends before a cooperative path becomes rationally conceivable. Agent **B** has a unique admissible resolution of the problem of choice depicted in Figure 5.4.4. Consequently, from the point of view of the present problem of choice, the best **A** can do is to pull down. This is exactly what the **BI** argument predicts.

The same holds for certain short-sequence Centipedes, as well. The situation illustrated in Figure 5.4.5 does not seem to offer an agent enough assurance of the prospects of trying to induce rational cooperation.

---

[177] Cf. especially Taylor (1987) and Tuomela (1992).

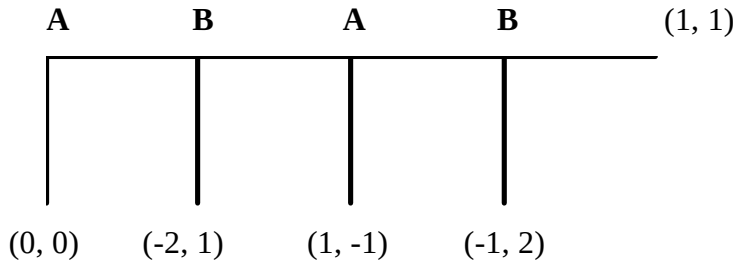


Figure 5.4.5 a short-sequence Centipede

Suppose that **A** finds himself at the first node in a situation as illustrated in Figure 5.4.5. Then it would be rational for **A** to move forward if **A** could expect **B** in turn to move forward, which is the case if **B** could rationally expect **A** to move forward at the following round(s). That is, rational cooperation on part of **A** depends on his prospects of inducing rational expectations to **B** about his future behavior. These expectations are in part induced by making the cooperative move, and thereby paving the way for a cooperative path, but also by the knowledge of the situation and the recursive capacities of the agents. Suppose that **A** can easily simulate **B**'s deliberations in the given situation. Then, **A** would see that it would be rational for **B** to move forward, if she could expect **A** in turn to move forward, which would be the case if **A** could rationally expect **B** to move forward during the following round (which happens to be the last round). Since **A** cannot rationally expect that **B** will move forward at the given node, then **B** cannot rationally expect **A** to move forward. Consequently, **A** would not succeed in inducing rational expectations about cooperative future behavior. In this case a cooperative move on part of **A** would not be a credible signal for **B**, since adequate prospects about the course of the game would be missing: the only mutually beneficial outcome (1, 1) is inaccessible. That is, the cooperative path is not rationally conceivable.

However, when the sequence of the Centipede is stretched further, the assurant properties may become visible, as Figure 5.4.6 shows.

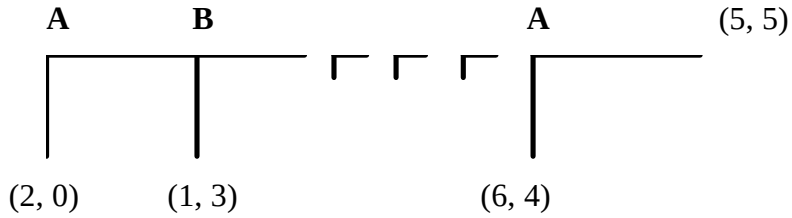


Fig. 5.4.6 the assurance problem & the Centipede

[178] They can, however, still be externally authorized.

The Figure 5.4.6 illustrates the very point of the assurance criterion from the point of view of agent **A**. By maximining, **A** can avoid the minimum of 1 and gain 2, but by doing so **A** would also lose the option of gaining more in the future. In this case the path that could be induced by moving forward promises more than timid maximining. If **A** could only focus on a bit more distant target than direct maximizing at the first node, and try to induce agent **B** to move forward on the expectation that **A** will also move forward, the assurance problem would become visible. Then, consequently the cooperative path would become rationally conceivable, since in an assurance problem cooperative moves are rationally conceivable. Figure 5.4.7 illustrates a situation in which the points of confidence (marked by the circles at nodes) are at a relatively early level in **A**'s deliberation process. It shows that if **A** can induce **B** to move forward, both **A** and **B** will gain at least the credit and even more than they would have gained were **A** to have pulled down at the first node.

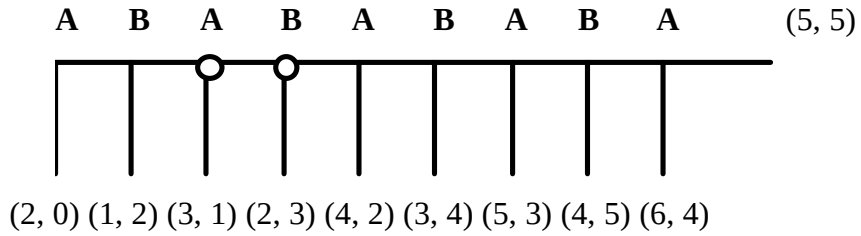


Fig. 5.4.7 the points of confidence

That is, from the point of view of the first nodes of **A** and **B**, respectively, the encircled nodes provide at least as much as the first nodes. So, the path that implements the options expressed by the encircled and the further nodes is rationally conceivable.

5.4.2 Salience

Condition (ii) of the (**IR**) introduces the idea of conformity in a preliminary way. It states that instead of full deliberation of all the relevant options, the present move is in part tied to one’s experience of the past cooperative events. Thus, rather than being based on an independent choice, the corresponding move is intertwined in the web of reciprocal expectations of discernible cooperative moves that in part determine the rational course of action in a given situation. Conforming to the pattern provided by the experienced cooperative behavior and reciprocal expectations of future cooperation shows a path out of

at least the basic problem of indeterminacy that looms in the structure(s) of the situation. In addition, it may also provide an exit from social dilemmas that exemplify the recurring problem of unsatisfactoriness.

Condition (ii) of the (**IR**) refers to the salience of the solution. An existing and recognized path itself is the salient option.[179] It is conformed to and every one expects the others to conform to it in part because it is a salient option, viz. a mutually known and existing solution to the present problem in recurring interaction. Furthermore, the path reproduces itself by the agents treading on it, and the salience of the resolution increases by this way. This idea is in perfect accordance with Lewis (1969): general conformity is strengthened by each token act of conformity.

It can be said that a central result of interpreting social habits in terms of self-enforceable paths is that salience becomes accounted for automatically. In order to resolve the problems of indeterminacy, social habits must be publicly observable. If they were not, the agents would be compelled to deliberate which of the available options to choose, and the point of the habit would be missed. There would not be a salient option to which to conform.

In analogy with a path being trodden, social habits are strengthened by each token act of conformity. The idea is the same, even though the awareness of public observability in case of habits may be more indirect than treading a physical path. Sometimes observing overt behavior may suffice for spotting an existing habit and sometimes it may require recognition of a sanction that regulates the corresponding behavior. Either way, this is in perfect accordance with Lewis' view. For Lewis' definition points out the publicly observable feature of the social habit: it is the regularity ***R*** in the behavior of members of a population ***P*** that exists and is recognized by the members of the community, and all this is assumed to be common knowledge within the members of the corresponding community.

5.5 Funk Games

An important feature of applying social habits to solve the problems of indeterminacy, of facilitating social interaction and, thus, emphasizing conformity as the move that implements the social habit is that standard game theoretical notions are put aside. According to the present account, then, conventions and norms are not seen strictly speaking as an implementation of *strategic choices* that amount to equilibrium. Instead, a notion of *conformity* to a rationalizable path is invoked. This feature of the account should be evident when we observe the definition of social habits (**SH**). Recall that:

[179] I.e. it is true, and common knowledge, that there is such a path.

(SH) a regularity **R** in the behavior of members of a population **P** when they are agents in a recurrent situation **S** is a *social habit* if and only if it is true and is common knowledge in **P** that
(1) virtually everyone prefers to conform to **R** on the condition that the others do, since **S** represents a problem of indeterminacy, and uniform conformity to **R** provides a rationalizable path out of the problem represented by **S**;
(2) in part because of (1) virtually everyone *expects* everyone else to conform to **R**; and
(3) in part because of (2) virtually everyone *conforms* to **R**.

The argument for adopting condition (1) was presented in the two previous sections. However, its status requires clarification. When it is a question of full conformity and corresponding expectations, the agents are unhesitatingly making a routine move. Although condition (1) in part gives a reason for making the expected move, this reason may be, and, in conforming, is not articulated. In a case of an arbitrary agent it may provide a partial reason for an action without ever being reconsidered or brought to the surface by means of a rational deliberation process. But it is nevertheless there. Furthermore, it makes the expected course of action salient under the given circumstances. So, what is left at the surface is conforming to a mutually expected course of action that implements the social habit.

5.5.1 Rational Behavior without Present Deliberation

The notion of conformity is crucial for introducing funk games. It resists the idea of making a deliberated choice in favor of moving along on the salient path. In collective action it is easier and safer to act in accordance with generally accepted customs than to try unexpected and unconventional but perhaps well deliberated alternative options. In fact, the rational option is generally the one that is in accordance with the mainstream. Then, it is also unnecessary to deliberate over and over again which is the best alternative under the given recurring circumstances. In contrast, deliberating the best choice in complex recurring interaction situations may turn out to be a wasteful expenditure of capacity and thus, in a sense, an inefficient attempt to seek the best alternative. In short, in the case of ordinary agents spending deliberative capacity may turn out to be unreasonable when a generally accepted option is available. This is especially true when the generally accepted option carries along the memory of rational deliberation in the past. Rationality is, in a sense, conserved in the practice. Furthermore, the existence of a salient social practice provides the agents with blinkers that, in a sense, inhibit reconsiderations of the reasonability of conforming. Alternative options are rarely observed when

routine moves are performed.

However, even though the idea of deliberating on all the available options is resisted, the basic notion of rationality as an optimal use of one's resources and skills retains its force. Rational agents are still seen as e.g. the maximizers of expected utility, although more indirectly so. Besides making a deliberated choice that ends up in equilibrium, rational action can also, and more readily so, be seen as moving along a path without present deliberation and merely as conformity to that path without reconsidering any other options.[180]

Under the given description, then, the basic game theoretical view of rational behavior in social interaction is no longer valid. At least it is inadequate. For this reason I introduce a novel class of games that are in part based on a rational reconstruction familiar from the standard game theory. But these games are also based on taking into account the connection of experience and the previous deliberations in what is considered to be the rational option under the given circumstances. This means, especially, that the model in question is a functionalist one. That is, there is a certain connection between the world and the practices of ordinary agents. For instance, the rationality of a move is not detected merely in terms of the plain preferences alone, but in part in terms of the broad circumstances, as well. The path, in part, determines what the rational move is under the given circumstances. Hence, the corresponding situations are better described in terms of a novel class of games which I have chosen to call funk games. In particular, funk games are described as games in which agents conform to, and are mutually expected to conform to, an existing and commonly known path.

Funk games can be said to be offspring of folk rationality, or rationality of ordinary agents, since in an ideal world the perfect skills of ideal agents would make the introduction of capacity-saving measures redundant. There is no need to conform to a habit, since all things are considered in deliberating of the present move, and all situations are reduced to their normal form, in which the external precedents have no meaning for ideally rational deliberators. Still, the behavior of agents in funk games falls under the description of rational interaction e.g. maximization of expected utilities under social settings. Only now, the rationality is brought about *indirectly* with respect to maximizing one's expected utilities. That is, the agents are not presently deliberating over the available options, as is the case with *direct* maximization of expected utilities, but are instead conforming to a known regularity. Under the description of funk games, the agents are acting for different reasons than under the description of standard games. In standard games the reason for action is something that falls under the description of direct maximization of expected utilities, whereas in the case of funk games the

[180] Cf. Bratman (1987) for a similar idea of personal policies.

reason is satisfying mutual expectations. This behavior, however, can be rationalized in terms of maximizing the expected utilities, but this does not give a correct description about what actually happens for a reason when an agent conforms to the known regularity. In funk games the maximization of expected utilities is the function of satisfying mutual expectations, and it takes place by conforming to a prevailing practice.

The emergence of funk games requires a substantial degree of coordination between the agents. This is due to basing funk games in part on the signal account and to the fact that they are expressed by means of e.g. social habits.[181] The agents must have had a sufficiently long common history in order to appreciate each other's rationality and they must have come up with a resolution to the problem of indeterminacy. By this I do not mean to say that the agents must know each other personally, but that their beliefs must have been induced by the same or similar (external) space of events and experiences that an existing path points out. For instance, the corresponding agents may have gone through the same social system in order to have common expectations in certain situations. After all, it can be taken as a matter of fact that conventions, norms and institutions are different in different cultures.[182] Ponssard (1990, p. 82) speaks of the theory of collective rationality: there is no point in being rational alone. Especially, there is no point in stretching the recursions of beliefs any further than ordinary capacity allows. The agents wish to keep their reputation of being rational and appreciating each other's rationality when interacting with each other under dynamic circumstances. This means that rather than constantly deliberating different options and making choices according to the deliberation process, the signal account suggests that the agents conform to certain induced patterns. Especially, the path itself in part determines the rational thing to do. But this also means that rationality itself seems to be in part a matter of a social habit that may differ from culture to culture, since when the process of deliberation is not extended further than the boundaries of ordinary capacity, the corresponding habits may drift to different paths. Furthermore, then, the concept of choice seems to lose some of its applicability in favor of conformity due to viable expectations.

Since funk games resist present deliberation and making a strategic choice, one might conclude that we are, in fact, dealing with an evolutionary account of games.[183] However, it must be emphasized

[181] In contrast, the standard games are expressed in terms of strategy *choices*.

[182] This does not mean, however, that customs of one culture cannot be learned by a member of another culture, and it certainly does not mean that an infant or a newcomer cannot learn the prevailing morals.

[183] Evolutionary games are, roughly put, games in which the agents are acting as if they were playing certain permanent roles, as e.g. HAWKs and DOVEs in the game of **H/D**. That is, the agents in evolutionary games do not choose between alternatives, but instead they implement a type. Thus, HAWKs always play noncooperatively and DOVEs always play cooperatively in situations illustrated in the following figure:

	D	**H**
D	(3, 3)	(2, 4)

that despite these similarities, funk games do not represent an evolutionary account. They still belong to the class of eductive analyses, although their moves are not the result of present deliberation. The moves of the funk games are still based on reciprocal inducement of rationally conceivable path, although the initial process of deliberation may be and is covered by conformity to a prevailing habit.

A main distinguishing feature between the accounts is that while the agents in an evolutionary account never act against their type, in funk games rational agents may slip to a path of crime, so to speak. The agents may act against social habits in order to try for instant maximization of expected utility despite the general expectation of conformity. This is, for instance, the very point of free riding, which is also adequately accounted for in terms of funk games. It must be emphasized that free riding is based in part on knowledge of the non-deliberation of the agents that conform to the social habit. The reverse of this possibility is that general conformity cannot be ultimately guaranteed. That is in part because the agents are essentially fallible. The essential incompleteness of rationality applies also to collective rationality.

5.5.2 Functionality of Funk Games

Instead of giving an explanation in terms of presently deliberated choices of fully conscious ideally rational agents, funk games, in a sense, provide explanations partly in terms of functions of the patterns of the agents. Lewis' (1969) account of convention can be considered a typical illustration of funk games in action, since instead of choices it speaks of conforming to a certain pattern or regularity in behavior. Let us speculate a bit with this idea in the light of Elster's (1983) definition of functional explanation in the social sciences.

In the funk games the function of conforming to a social habit is *to maximize expected utility* under the circumstances of social interaction, viz. a game. However, since deliberation over choices is not directly at stake, the maximization of *expected* utility is not direct. Thus, agents do not, strictly speaking, make a choice, but instead they conform to e.g. a cooperative pattern, social convention or prevailing custom. The conforming agents are not conforming because this is the way they maximize their expected utilities but because this is the way they are expected to act under given circumstances. However, by mutually conforming, the expectations of the agents are fulfilled. The interaction between the agents is facilitated and their deliberational capacities are reserved for situations of true need. Furthermore, the *status quo* of the social circumstances is sustained.

H (4, 2) (1, 1)

As long as the underlying sequential game goes on and its end is beyond the horizon of deliberational capacity, the funk game may continue, and the pattern serves its function. However, as soon as the inducement of the corresponding rational expectations according to the (**IR**) recommends switching to the deliberation according to the standard requirements, the function ceases to exist, the pattern is no longer valid as such, and the funk game may dissolve, since it becomes *dysfunctional*.

The funk game succeeds in providing a true functional explanation if it is in accordance e.g. with Elster's (1983) characterization of functional explanations in the social sciences:

An institution or a behavioral pattern X is explained by its function Y for group Z if and only if:

(1) *Y* is an *effect* of *X*;

(2) *Y* is *beneficial* for *Z*;

(3) *Y* is *unintended* by the actors producing *X*;

(4) *Y* (or at least the causal relationship between *X* and *Y*) is [virtually][184]*unrecognized* by the actors in *Z*;

(5) *Y* maintains *X* by a causal *feedback* loop passing through *Z*. (Elster 1983, p. 57).

Jack Vromen (1995, p. 99) approaches Elster's definition as follows:

> The first condition speaks for itself. To say that *Y* is a function of *X* is to say that *Y* is caused by *X*. The second condition also seems to be implied by saying that *Y* is a function of *X*. Harmful or deleterious effects would not be called functions. Condition (3) and (4) make clear that *Y* must be a latent function. It is possible that the actors who produce *X* do so because they are striving after some goal. But their goal may not be *Y*. Otherwise; *Y* would be a manifest function. And the explanation would be intentional, not functional. The same holds for recognition. As soon as the actors involved recognize that doing *X* causes *Y*, *Y* ceases to be a latent function. Finally, condition (5) demands that a causal feedback loop, running from *Y* back to *X*, must be demonstrated.

The following can be said to hold for funk games: instead of making a deliberated choice, conforming to a social habit (*X*) produces as its effect the maximization of expected utility (*Y*). This non-deliberational feature can be said to be unintended by those who conform to the social habit. The intended goal is to act as expected, e.g. obey a norm. Not deliberating is naturally beneficial for the group (*Z*), since it guarantees continuing conformity to the pattern that produces satisfactory outcomes. Thus, funk games are in accordance with the conditions (1) - (3). What about conditions (4) and (5)? I added [virtually] to condition (4), since during the act of conforming to a habit, the agents do not typically recognize its function as stated above. They do not contemplate at the time of conforming whether they are at that moment serving the function of facilitating interaction etc., nor is this typically a reason for them to act accordingly. The reproduction of the pattern, on the other hand, takes place by e.g. appealing to the

---

For a more detailed discussion cf. e.g. Sugden (1986), Young (1993) and Vromen (1998).

[184][Virtually] added by me.

reciprocal interdependent expectations about the pattern and its manifest benefits, not by appealing to the function itself, which is, as such, irrelevant and arbitrary for the agents (other than social scientists). Finally, when the nature of funk games is correctly understood, condition (5) speaks for itself. As long as the status quo circumstances provided by the continuous maximization of utilities (*Y*) are feasible for the group (*Z*), that is, as long as the members of the group (*Z*) succeed in satisfying their expectations, they continue conforming to (*X*) in the non-deliberative and capacity-saving fashion, and *Y* can be said to maintain *X* due to the referred causal feedback loop through *Z*. Thus, a cooperative pattern, a convention, an institution or a norm is reproduced by *de re* recognizing that the right circumstances for running a backward induction do not prevail, and in order to manage as well as possible under the circumstances, one is better off complying with the prevailing pattern. This compliance is beneficial for the group or collective *as a whole*, as stated in condition (3). But if it would turn out that this compliance were strictly inadmissible for the actors producing the cooperative pattern, they would turn their back on it no matter how beneficial its effects were for the collective. Consequently, as soon as the members of the group are in a factual position to run the backward induction, the circumstances are altered from those that can maintain the continuance of the pattern, and the functional basis for the reproduction of the pattern ceases to prevail. The collapse of the functional basis does not, however, always lead abandoning the action that is in accordance with the social habit; it only means the agents succeed in motivating these acts of cooperation from another basis than that ultimately based on rational deliberation.

I do not wish to go deeper into the discussion of the functional explanation here. The purpose of the above analogy was to motivate the notion of the funk game in a preliminary way. My aim is to introduce and define the notion of the funk game and propose that it should replace e.g. Lewis' notion of convention in discussions of the emergence of cooperation in loosely game theoretical terms. A simple reason for this is its utmost generality when compared to e.g. the notion of social convention. Besides covering merely conventions and the coordination problems, funk games cover norms, institutions and a vast range of social problems from which the norms and institutions emerge, e.g. the problems of cooperation.

It is, however, instructive to begin the introduction of funk games by investigating the social conventions. I take it that Lewis' account in many ways anticipated the signal account and, consequently, the funk games introduced in this essay. Not only did Lewis introduce how agents conform to a convention that dictates what the agents are expected to do, but he introduced the idea of *higher-order expectations* defined by recursion (1969, p. 28). He also emphasized the idea that from the *regularity* of past cases it is reasonable to extrapolate to the near future, and that this regularity adds to

the agents' experience of *general conformity* (1969, p. 41). This is the very essence of the signal account, as stated in the previous chapter. Lewis considered only the solutions to the coordination problems, but as the case of the emergence of cooperation shows, the solution rests on the possibility of making certain generalizations concerning the recursions of expectations, and in this way of inducing reciprocal expectations on cooperation as well. This possibility of making the generalization is what is, however, missing from the signal account, and this is the distinguishing factor of funk games.

6 FREE RIDING AGAIN

Let’s pull the loose ends together. In Chapter 2 the problems of rational collective action were the main concern. It turns out that the standard game theoretical approach cannot account for full-blown rational collective action. In its basic form the problem of rational collective action is seen as a conflict between individual and collective rationality with the consequence that rational collective action never occurs: if the two seem to conflict with each other, the winner is individual rationality. This is due to the precondition that rational collective action is nothing more than a mereological combination of individually rational actions - and then the apparent conflict is based on a misconception of rationality. This trait of accounting for rational collective action makes collective rationality notoriously epiphenomenal.

In order to account for collective rationality as a genuinely effective motivational factor, the aspect of coordination is, according to an alternative approach, thought to come to the rescue. Changing the focus into coordination does not undermine the efficacy of individual rationality, but it brings along certain novel properties that cannot be captured in terms of e.g. **PD**-type problems alone. Mutual expectations and coinciding interests come into focus when the question of coordination is at stake. Then, achieving collective rationality is ultimately a question of an interdependent choice. However, the standard game theoretical approach does not give any clue how the interdependent choice could be implemented. This is thought to be an inevitable consequence of the causal independence of agents. However, as e.g. the analysis that goes in terms of the correlated equilibria shows, causal

independence does not rule out probabilistic interdependence.[185] For instance, the agents may bind their actions together *in beliefs* and induce expectations about certain behavior when the conditions are suitable.

The basic problem of making an interdependent choice, however, is the possibility of indeterminacy, which is not resolvable in terms of perfect equilibrium alone, as the vast literature on the subject testifies.[186] Binding actions together *in beliefs* does not ensure the necessary *stability* for collective rationality to prosper. As I claimed in the previous chapter, the notion of equilibrium does not succeed in giving an adequate description of the mechanisms underlying rational collective action, and its basic problems. The focus must be shifted from deliberating on alternative options taking into account the social practices that facilitate interaction between rational agents. These practices put the emphasis on conforming and expectations of conforming in contrast to recurrently reconsidering whether to go this or that way, and what the neighbor will do next, etc. Especially, the stability required for collective rationality is gained by means of conformity and reciprocal expectations of this conformity. Thus, stability is explicated in dynamic terms of recurring expected actions rather than in terms of equilibria that occur in beliefs alone. An existing path of recurring actions amounts to salience that works to enforce the path itself in terms of conformity. This is what the stability of collective rationality is all about. No present deliberation is needed. Instead, the moves are made in a routine fashion. Consequently, the standard game theoretical description of rational behavior no longer applies. The conditions of collective rationality are better described in terms of funk games. In short, this means that the emphasis is put on the notion of conformity instead of the notion of choice.

However, the standpoint of making the present deliberation factually redundant cannot be motivated by assuming ideally rational agents. Ideal agents would have no need to depend on social habits, since their skills and capacity of making the necessary recursions do not run into any boundaries. Hence, it would be pointless to try to induce rational expectations for cooperation in finite dilemmas. Furthermore, it would be pointless to speak about conformity to a path, since there are no limitations to being able to weigh all the available options. Furthermore, in case of the problem of indeterminacy, the ideally rational agents are happy with the result of randomizing the choice, since they are happy with the result of backward induction in the corresponding situation. Ideally rational agents do not have any qualms about the counterintuitivity of certain results as long as those results are in accordance with the

[185] Cf. Aumann (1987) for discussion.
[186] A simple example of this result is provided by Gilbert (1990).

principles of coherence and consistency. So, the standpoint is that of ordinary agents with imperfect skills and limited capacities.[187]

Counterintuitivity appears as a problem of rationality for ordinary agents when, for instance, their limited capacity does not suffice for making the necessary recursions for backward induction. Then, it is possible to look forward and realize that the prospects of inducing cooperation look promising. If one agent can expect another agent to think what he is thinking, then inducing cooperation appears to be a rational option. A similar mechanism can be applied in resolving the problem of indeterminacy, for instance in situations of recurring problems of coordination.

Optimal use of one's resources does not require recurring deliberation in recurring situations. This is especially true when the capacity to deliberate is one of those resources, and these resources are limited. Then it is rational to adopt habits that facilitate e.g. social interaction in the sense of conformity to an expected course of action and avoidance of reconsidering alternative options. Making use of previous deliberations and tying one's actions into the reciprocal expectations in this way means that the optimal use of resources is gained in a collective way. That is, collective rationality gives reasons for actions in recurring social situations.

***

The questions of collective rationality and collective action are especially perplexing in accounting for free riding.[188] The reason for reconsidering free riding lies in its special ability to deal with the circumstances of coexistence of collectively and individually rational behavior. In order to account for individually rational and collectively rational behavior in coexistence, two distinct sources of rational collective action need to be taken into account. Roughly put, the sources are the problem of cooperation and the problem of coordination. The former is meant to represent the problems of unsatisfactoriness and the latter the problems of indeterminacy. These problems represent the ultimate basis for the problems of collective action, but they need to be explicated in conjunction if free riding is to be accounted for at all.[189] In their distinct forms the corresponding problems of rational collective action are often illustrated in terms of the **PD** and the **CC**. Separately, and in their normal forms, however, these problems give incomplete picture of free riding. The problem of free riding does not

[187] Cf. the discussion in Chapters 3 and 4 of this essay.
[188] Recall the discussion in Chapter 2.
[189] Cf. the discussion in Tuomela (1992) and Hampton (1987). They and other authors point out the elements of coordination in explicating free riding. However, their accounts are incomplete from the point of view of the present essay.

reduce to either of these models. Nor does it reduce to any model that can be illustrated in terms of a simple normal-form game structure, since the standard solution concept of game theory does not succeed in capturing the problem.

Free riding is in part based on the deliberation of alternative options nesting in a situation of general conformity to a certain collectively accepted and mutually expected social pattern. Conforming to this pattern means, especially, non-reconsideration of the salient option and non-deliberation of the alternative options. Thus, funk games provide a fertile ground for accounting for free riding, or so I claim. However, a crucial question is still open. Under circumstances of general non-reconsideration and non-deliberation it is hard to figure out how one comes to think of the option of taking a free ride in the first place. It seems that the (**SH**) underlying the funk games does not give any room for making an unconventional move instead of conforming to the salient option. A possible answer may be found when a *free ride effect* is realized. That is, the possibility of benefitting from the efforts of others without one's own contribution. This effect is typical of collective goods.[190]

In general, the free ride effect is involved in the resolution of the problems of unsatisfactoriness. Especially, this resolution is based on giving up the postulate of uniqueness in order to enforce a result that is in accordance with folk intuitions in recurring situations. In the long run, the cooperative option path may become salient in the sense that the defective option is no longer considered. That the alternative option is in general no longer considered does not mean that the salient option has become the uniquely best option, although the salient option is the one that virtually everyone conforms to. Things being thus, the general conformity leaves room for occasional present deliberation over alternative options without this undermining the social habit and the common knowledge of rationality. Especially, there is room for a free rider to squeeze in without this immediately deteriorating the corresponding social habit, since it can be generally expected that virtually everyone conforms and expects the others to conform to the prevailing practice.

The free ride effect may be evident for aliens and infants, for they might not have adopted the customs and conventions of our society. Fortunately, they are allowed to make mistakes without serious consequences until it is reasonable to hold that they must be well aware of the prevailing practices. But when full-fledged and adequately informed members of the community are concerned, many find the rationale of breaking the rules perplexing. This is due to the social blinkers provided by the prevailing customs. In particular, these blinkers have the effect of making the unconventional options factually ineligible. Even if the option of free riding were realized, the agents would tend to conform to the

[190] Economists might say public goods.

salient option under normal conditions. This result is dictated by the (**SH**). In contrast, free riding is an option that conflicts with the (**SH**), and confusion is at hand. The general conformity to the prevailing practices is what makes free riding possible, provided that it involves the free ride effect. However, general conformity resists the idea of reconsidering the prevailing practice. A crucial question is, then, how does free riding emerge under circumstances of general conformity. Under normal conditions the blinkers provided by social habits make the unconventional options unsalient. If virtually every one conforms in part because it is mutually expected and performs the conforming act without reconsidering or even noticing alternative options, then how does one come to think the option of free riding in the first place? How does one come to think that there can be and is such a phenomenon as the free ride effect, and that it can be applied to one's own benefit?

The confusion can be cleared up, though. Ordinary rational agents are innovative – they are capable of inducing rational expectations about cooperation when the basic options are unsatisfactory, and they are capable of learning and adopting habits that facilitate interaction when they face indeterminate solutions. This creativity does not disappear when the problems of unsatisfactoriness and indeterminacy are resolved. Especially, when in distress, people begin to search for novel solutions, and when given enough time, some may probably find some of those solutions. For instance, some one may discover that the free ride effect may be utilized without spending one's few resources on conformity when an immediate benefit is beyond reach.

The general resistance to unconventional options, or at least their general lack of salience, provides the required stability of collective action. Provided that the free ride effect is involved and recognized, an agent has a partial reason to reconsider the possibility of benefitting from the others' willingness to contribute. This is in part due to the fact that when the problems of unsatisfactoriness and indeterminacy are resolved by means of general conformity to social habits, it is virtually inevitable that there will be enough contributors. Furthermore, the general insalience of an option means, especially, that virtually everyone is disposed to believe that under normal conditions it is unnecessary to take the unconventional option into account. For if conditions are normal, an agent may arguably reason like the Individual in Hardin's (1982) example and expect to get a free ride without being guilty of committing the fallacy of composition.[191] Let us elaborate these observations presented above and in Chapters 2 -5 into an analysis of free riding.

[191] Cf. Chapter 2 of this essay for discussion.

6.1 Accounting for Free Riding

Under certain circumstances the pace of general conformity involves circumstances in which a free rider may strike without virtually any one noticing it, or at least not minding it. A free ride may go unnoticed, since it does not always involve a direct cost to those who conform, and sometimes interfering in a sudden free ride may be more costly than letting be. Even at the aggregate level the costs of preventing free riding may exceed the costs of tolerating it. In the standard approaches, however, there is no room for free riding. Either a conflict comes about, in which case each agent maximizes his expected utility by defecting, or a problem of coordination occurs, in which case the agents try to find a mutually expected resolution. Instead, funk games allow that despite the general conformity and the suitable expectations, some may go against the mainstream without this questioning the common knowledge of rationality and the rational conceivability of the cooperative pattern. Or, to put it alternatively, some may keep pace with the general conformity without contributing to the collective goal, without this questioning the rationality of those who contribute. If it did, then collective action would be notoriously unstable, and perhaps even inaccessible. As a matter of fact, the non-deliberative and non-reconsidered characteristics of conforming to a commonly accepted and acknowledged regularity, viz. a social habit, creates under certain circumstances the possibility of free riding. In general, this feature can be called the free ride effect.

Funk games are supposed to provide fertile ground for free riding. The (**SH**), presented in the previous chapter, lays down the conditions that must be satisfied before the eligible expectations for getting a free ride may come about. Granted that the (**SH**) holds and that the group in question has found a resolution to the problem of cooperation, e.g. a sequential **PD**, then from the point of view of an individual, the following must be true:

(ESH) Agent **A** may expect that a suitable social habit[192] (that accords to the (**SH**)) is being followed if and only if there is a prevailing cooperative pattern *X* such that

(1) **A** believes that virtually everyone prefers to conform to *X*, on the conditions that the others do, since conforming to *X* yields a rationally conceivable resolution to a social dilemma *S*;

(2) in part because of (1) **A** expects that virtually everyone expects everyone else to conform to *X*; and

(3) in part because of (2) **A** expects that virtually everyone conforms to *X*.

The (**ESH**) is entailed by the (**SH**) when the focus is on the standpoint of an individual member of the community. The (**ESH**) is meant to capture among other things an agent's expectations of the circumstances that ultimately are conducive to free riding. The prevailing cooperative pattern *X* refers to a course of action which involves the free ride effect, that is, for instance, a situation involving a sequential **PD**, but not a sequential **CC**. However, free riding is not mentioned in the above characterization, since the purpose at this point is to describe the conditions in which free riding may emerge, given that certain extra conditions are fulfilled. Especially, when the (**ESH**) holds, it can be expected that agent **A** will conform, unless additional, conflicting, reasons occur. The *awareness* of the free ride effect belongs among those reasons. But being aware of it is only one step in starting to deliberate over the less salient options and to reconsider the reasonability of conforming.

The reasonability of conforming may raise an issue if the free ride effect is realized. In the standard analyses of free riding this means, especially, that a rational agent realizes that there is a non-excludable good attainable by a group, and it can be attained by *K* members of the group, where *K* is less than all members.[193] However, just being aware of the effect is not as such sufficient for the motive of free riding. The *suction* of general conformity resists defection even if the free ride effect is realized.

A crucial element in the standard argument is to realize that when the amount of contributors falls below the level of *K*, the joint action opportunity of providing the common good ceases to prevail. That is, the corresponding group is assumed to be latent.[194] Thus, as Pettit (1986) expresses it:

"The fear of contributing when the good is not produced, and the hope of not contributing when it is, make it rational for each person not to contribute to the production of the good: specifically it means that the strategy of not contributing maximizes expected utility."

Obviously, in contrast to what Pettit and others claim, this condition does not qualify as a rational motivation for free riding, for it lacks the required stability of collective action, and it is in part in conflict with the (**ESH**) that guarantees the right circumstances for free riding. That is, it lacks the circumstances of a stable and continuing course of mutually expected cooperative action. Thus, the motivations for free riding need elaboration so that they can be accounted in terms of the funk games in which moves based on present deliberation are nesting.

Tuomela's analysis (1992) brings the basic features of free riding together. Tuomela's definition of intending to free ride is as follows:

---

[192] Note that a social habit may also emphasize coordination in the sense there is no free ride effect involved. For this reason I speak of suitable social habit on this occasion.
[193] Cf. e.g. Pettit's analysis, (1986), p. 367. Cf. also Hampton (1987).
[194] Cf. Pettit (1986), Hampton (1987), and Tuomela (1992) for discussion.

(FR) A member A of a collective G *intends to free ride* relative to a public good produced by a joint action X if and only if

(1) A intends to defect (viz. not to contribute or do his part of X).
(2) A has a belief to the effect that the joint action opportunities for the performance of X will obtain, especially that at least K members (or a sufficient number of members required for the provision of the public good produced by the performance of X) contribute (or do their parts).
(3) A believes that he ought to participate in the production of X and there is (or will be) a mutual belief among the full-fledged and adequately informed members of G to the effect that the joint action opportunities for the performance of X will obtain, and to the effect that each full-fledged and adequately informed member ought to contribute.
(4) A believes that he will gain more by defection than from contribution if at least K agents contribute, where K is the minimal numbers of agents capable of jointly performing X.
(5) A believes that the outcome resulting from all the agents contributing is better than the outcome when all defect.
(6) A believes that his defection involves a cost (possible nil) to the contributing members of G. (1992, p. 174)

Now this analysis can be accounted for in terms of funk games. Condition (2), (3) and (5) can be traced directly from the (**ESH**). Supposing that the (**ESH**) can be taken for granted, agent **A** has reason to believe that the joint action opportunities of performing **X** obtain, and that at least **K** will contribute, since it is true and mutually believed that there is a social practice that virtually everyone conforms to. In this case the prevailing practice is such that it provides a free ride effect, e.g. a norm of cooperating in the production of the collective good. Instead of deliberating over the possibility of all agents deliberating upon the available options, agent **A** can trust his move on the expectation that virtually everyone will conform to the practice. Thus, agent **A** may reasonably (in a *de re* sense) expect that virtually everyone is playing a funk game, since it is common knowledge that virtually everyone tends to conform instead of making a deliberated choice. This does not mean that agent **A** could not be mistaken in his expectations. After all, **A** is fallible. Furthermore, the belief in rationality of conforming may not be articulated in any way. Conformity is just what is in general expected on the part of the other agents, and it is not very plausible to expect that the prevailing practice would dissolve just like that.

Condition (3) of (**FR**) refers directly to the mutual expectation that all the full-fledged and adequately informed members of the community, including agent **A**, will conform. In this case it means doing one's part as expected and not reconsidering other options. Condition (3) of (**FR**) gives a partial reason to the effect that (2).

Condition (5) of (**FR**) is supposed to give a partial reason to the effect that (3), and, consequently, to the effect that (2). However, it needs to be reformulated in order to yield the required partial reason. A mere belief that something *Y* is the case does not give reason to expect that *X* will be or ought to be performed, where *Y* will result from performing *X*. The performance of *X* requires making it rationally conceivable, which cannot be guaranteed just by referring to the virtuousness of the outcome. After all, *X* may be inadmissible e.g. due to the dominance of an alternative option. In that case, agent **A** would lack any reason to expect that others will contribute, as well as he would lack any

reason to expect the mutual expectation of the general conformity to performing *X*. Fortunately, I have presented an argument to the effect of making the cooperative option rationally conceivable so that the partial reason to the effect that (2) and (3) can be held.[195] In short, condition (5) of (**FR**) should state that conforming to the performance of *X* resolves a social dilemma, e.g. producing a collective good by means of cooperating. In addition, in part because of its ability to resolve the dilemma, the performance of *X* is preferred by virtually everyone. It is worth noting that if performing *X* is strictly dominated by another option, it cannot produce the required resolution.[196] Furthermore, the partial reasons are provided by means of conformity, which, in turn, is better described in terms of following a rationalizable path rather than making a deliberated choice concerning all the relevant or possible options. In the case of the possibility of free riding, this would mean conformity to a cooperative pattern that has proved to yield more promising results in recurring social situations than the option of direct maximization of expected utilities.

Now, conditions (2), (3) and (5) of (**FR**) consist of the body of the (**ESH**). However, conditions (2) and (3) are the more crucial ones, since (5) may not be articulated in the sense that the deep reasons for adopting e.g. a cooperative habit are no longer reconsidered. Instead, cooperation may take place as a routine move that accords to the mutual expectations about the current behavior. That is, people contribute because they are expected to contribute, and they do not deliberate further upon the reasonability of conforming to the expected pattern.

Conditions (2), (3) and (5) in part create the basis for the possibility to expect to get a free ride, but they do not suffice as such. Consequently, the (**ESH**) does not suffice as such. The circumstances must be suitable. Especially, performing *X* must produce a free ride effect, that is, it is possible to enjoy the results of cooperative activity without one's own contribution. In general, the free ride effect occurs when the resolution to the problem of unsatisfactoriness is at stake. When ordinary agents succeed in inducing rational expectations of cooperation, it does not take place by pointing out a unique solution to a problem of choice, but by introducing a path that promises a resolution to the problem of unsatisfactory outcomes. The path reinforces itself by reciprocal acts of cooperation. However, cooperation will never be a unique option in the course of action, although it may become the salient course of action, not least because it provides the resolution of certain social dilemmas. Although reciprocal cooperation in recurring situations may be considered a rationally conceivable and prospective option, an alternative option remains that might have vanished (because of the general conformity to the cooperative pattern) and that can be considered equally rational. It is the option of

[195] Cf. Chapter 4 for the presentation of the (**IR**) and Chapter 5 for the analysis of the (**SH**).

taking the advantage of the others' conformity without one's own contribution. This option is available when collective action involves a free ride effect – a possibility to benefit from the efforts of others.

This is probably what Tuomela (1992) may – at least in part - have had in mind when stating condition (4) of (**FR**). This means that agent **A** must be aware of a prevailing free ride effect, or at least believe that there is such an effect in general, and that there is a possibility to abuse this effect by not contributing to the performance of *X*. Especially, agent **A** must believe that he can free himself from the cost of contribution and still enjoy from the result of the others performing *X*, provided that the joint action opportunity prevails. This prevails if the (**ESH**) is taken for granted. However, taking for granted that condition (2), (3) and (5) of (**FR**), and especially the body of the (**ESH**), hold, the awareness of the free ride effect, as stated in condition (4), does not have to lead to any actions on the part of agent **A**, or of others. After all, it is quite reasonable to conform to the cooperative pattern, although defection provides instant satisfaction by maximizing expected utilities in the short run. That is because the defection may be fatal to the continuation of cooperative pattern in the future rounds, unless the defection goes unnoticed on the part of those who contribute.

Typically, no one will ask whether some one is getting a free ride if the defection does not involve a direct cost to the contributing members of the collective, and, typically, free riding does not directly add to the burden carried by the contributors. This is one of the characteristics of the free ride effect. An agent may keep pace with the general conformity without contributing himself to it while no one notice whether an agent enjoys a free ride. This is because the free ride effect gives the opportunity to allow those who conform to the cooperative pattern to bear the burden by the effort they would have made in any case. The cost involved by a free ride is indirect. It does not consume the resources of the contributing members as such, but as it does detract from the common efforts. It can be pointed out at the aggregate level only and by drawing a parallel to an idealized situation in which no one is benefitting from the free ride effect. The cost caused by a free ride can then be defined by comparing the amount of effort with and without free riding. In any case, from the point of view of being aware of the free ride effect, agent **A** may come to believe that his defection does not involve any relevant cost, or a cost in any relevant sense to the contributing members of the collective.[197] These observations refute condition

[196] Cf. the single-shot **PD**.

[197] The idea of an indirect cost seems to require clarification. An example from cycling does the job. Cycling is teamwork *par excellence*. It is typical that each cyclist in turn does the job at the peak and the others may enjoy the benefits of drafting. Those who follow the lead do not directly involve a cost, in the sense of consuming resources, to the peak cyclist. In general, taking turns at the head works pretty well. In fact, the free ride effect provided by rotating leaders is used to keep up the speed of the entire group of cyclists. The speed is not reducing, although some, and in fact most, of the others never take their turn at the peak position. They, in a sense, take a free ride, although they do not directly increase the burden of those who take their turn at the peak position. The peak cyclists set the pace, and the others follow and take

(6) of (**FR**) and suggest an alternative condition (6'): **A** believes that his defection does not involve any relevant cost to the contributing members of **G**. Together, condition (6') and (4) are supposed to provide a partial reason for (1) of (**FR**): **A** intends to defect.

However, I am not sure about this result. The pressure of general conformity may turn agent **A**'s intention upside down despite his awareness of the free ride effect and the possibility of abusing it. As a matter of fact, this question should be left open, since in discussing collective rationality there are no determinate answers. For one thing, the resolution is ultimately underdetermined when social habits are at stake, and for another thing, the possibility of free riding is based on an existing path of mutually expected conformity. But even so, the analysis tells something about an agent's belief or expectation regarding the possibility of getting a free ride. Accordingly, let us *reorganize* the above discussion and define the expectation of an agent's possibility of succeeding in free riding as follows:

(EFR) Agent **A** may expect to get a free ride if and only if there is a prevailing cooperative pattern *X* such that
(1) **A** believes that virtually everyone prefers to conform to *X*, on the condition that the others do, since conforming to *X* yields a rationally conceivable resolution to a social dilemma *S*;
(2) in part because of (1) **A** expects that virtually everyone expects everyone else to conform to *X*;
(3) in part because of (2) **A** expects that virtually everyone conforms to *X*;
(4) **A** believes that the general conformity to *X* involves a free ride effect;
(5) in part because of (4) **A** believes defecting from *X* involves no appreciable cost to those who conform to *X*; and
(6) in part because of (5) **A** believes that it is possible to abuse the free ride effect.

In short, free riding means the abuse of the free ride effect produced by the cooperative pattern *X*. The abuse of the free ride effect refers to the fact that as a full-fledged and adequately informed member of the collective, **A** is expected to perform his part of *X*, and **A** knows this. The same holds for all the full-fledged and adequately informed members of the collective, as well. The abuse also suggests that there is a correct use of the free ride effect, as well. In fact, the flip side of the free ride effect is that it is a basic feature of functioning collective rationality. I shall clarify this point in the following section.

Why free ride? Free riding yields sure-thing benefits when the (**EFR**) holds, even if it is based on shortsighted reasoning. Why conform? Even though free riding involves no direct cost to those who

---

turns in taking advantage of the leader's drafting. All cyclists in general get to enjoy this free ride effect, even if some

maintain the free ride effect, in the long run increasing free riding erodes general conformity. Ultimately, it is no longer reasonable to expect that virtually everyone will conform – or even that a required amount of the members of community will conform – in which case the expectation of getting a free ride is no longer valid. The answer to these two questions remains indeterminate. Ultimately, defections reduce the possibility of using the free ride effect for one's own purposes, although when there is such an effect it certainly will be abused, as well.

6.2 An Unexpected Virtue of the Free Ride Effect

It is important to note that the free ride effect provided by the *suction* of general conformity is not always corrupting from the point of view of the community or of even a contributing member of the community. The flip side of free ride effect is that not everyone needs to be active in order for collective rationality to prosper. The beneficial traits of the free ride effect are evident and, in fact, widely applied in Western societies.[198] The (**ESH**) supplemented with the expectation (or awareness) of the free ride effect provides a social safety net in case agent **A** is incapable of or for a legitimate reason is not doing his part of *X*. Agent **A** may trust the general conformity to prevail even if some are not presently contributing to it. An agent's expectation of being able to enjoy the benefits of the others' work can be expressed as follows:

(EBS) Agent **A** may expect to enjoy the benefits of belonging to a community *C* if and only if there is a prevailing cooperative pattern *X* such that

(1) **A** believes that virtually everyone prefers to conform to *X*, on the condition that the others do, since conforming to *X* yields a rationally conceivable resolution to a social dilemma *S*;

(2) in part because of (1) **A** expects that virtually everyone expects everyone else to conform to *X*;

(3) in part because of (2) **A** expects that virtually everyone conforms to *X*;

(4) **A** believes that the general conformity to *X* involves a free ride effect;

(5) in part because of (4) **A** believes that one's contribution is not always necessary for the general conformity to *X* to prevail; and

(6) in part because of (5) **A** believes that one occasionally has an opportunity to rest, educate oneself, or retire without it eroding the general conformity to *X*.

---

cyclists never contribute to providing it.

[198] E.g. in the welfare state, social services, pension benefits, sick leaves, being between jobs etc. Even annual holidays, culture and fine arts, science etc. can be said to be based on the free ride effect.

It is obvious that the act of free riding can be fitted into this description, as well. Ultimately, then, the difference between free riding and enjoying the benefits of a community is a matter of legitimacy. It is a question between enjoying and abusing the free ride effect. But that is an issue that would require a study of its own.

The (**ESB**) expresses the idea that can be called *the effectiveness of the whole*. In this case, the free ride effect can be utilized as a benefit to the society, instead of treating it a symptom of malaise. Especially, the free ride effect does not always require actions that aim at reducing it. However, the question of attitudes toward the free ride effect is a matter of political philosophy, and is thus the subject of another essay. Furthermore, the free ride effect and its beneficial effects cannot be explicated in terms of individual rationality alone. Conforming to the prevailing practices may be effective in the sense that people may trust certain conditions to persist even if one is not contributing or cannot directly contribute to their persistence. The free ride effect makes it possible that every one may get a turn to rest, or, at least, not to contribute directly, but e.g. to educate oneself. The effectiveness of the whole contributes to the persistence of the social structure by means of the free ride effect in that it is not necessary to exhaust the agents by demanding of them the full use of their capacities. The effectiveness of the whole puts the notion of optimality in a new light. Optimal use of resources does not mean maximum use. Among the reasons for adopting social habits is that it holds the capacity of making inferences in reserve until the real need arises. This may entail reconsidering social habits, as well. Even so, society is not a one-trip project, and optimal use of resources does not mean wearing them out. If the free ride effect could not be tolerated, then the present society as we know it would hardly exist.

6.3 A Final Comment

What do funk games add on top of all this? The answer is obvious. Funk games are games that are in part built on the standard games of game theory in that they are based on the core notion of rationality, the maximization of expected utility, and deliberations upon the best strategy for implementing the given goal. However, funk games emerge under circumstances of recurring social interaction situations and incomplete skills and capacities of the agents. Under these circumstances, then, in order to avoid expending limited capacity it is reasonable to adopt certain habits that facilitate social interaction while

still aiming at the maximization of the expected utility. Occasionally, individual endeavours, however, end up with unsatisfactory results, or the goal of maximizing the expected utility is inaccessible by the means of individual deliberations alone. Under these circumstances, then, in order to guarantee some security in their endeavours and to reach a satisfactory result altogether, it is reasonable to adopt certain social rules and habits that do the job. However, following such rules and conforming to such habits is no longer a matter of present rational deliberation, but a matter of compliance. Instead of choosing between alternative options, the agents are expected to conform to a prevailing practice. Thus, the standard game theoretical approach is no longer adequate.

We need an account that embraces the initial requirements of rationality while taking into account the extraordinary social properties of the resolutions of the initial problems of unsatisfactoriness and indeterminacy. Funk games deliver the goods. They are expressed in terms of social habits, such as social norms, institutions, customs and morals. The common denominator of all these social habits is that instead of deliberating upon all the alternative options, agents are supposed to conform to a prevailing course of action. In short, the term ‘funk game’ refers to the type of games that the social habits belong to.

However, social habits as such do not adequately describe all the properties of funk games. Social habits do, for instance, capture free riding even with respect to those social habits that have emerged to solve the problems of cooperation. This is due to the fact that social habits are a matter of conformity, but free riding is essentially based on present deliberation. Granted, social habits express the basic idea of the funk games, but do not account for the fact that rules can be broken. For this task we need a notion that covers a larger area than social habits as such are capable of doing. Such a notion is that of the funk game. Besides describing a situation in which conformity to a prevailing path plays the main role, it permits rational agents to occasionally consider alternative options even when they are expected to conform to the habit. This feature is implemented throughout the system of social habits by the fact that funk games are in part grounded on the standard notion of rationality and that the solution to the problem of cooperation under the given circumstances cannot be uniquely determined. Hence, even if general conformity has emerged in an arbitrary society, there will be those who go against the mainstream. The solution to this problem at the societal level is, however, a matter of another study.

REFERENCES

AL-NAJJAR N. (1995): 'A Theory of Forward Induction in Finitely Repeated Games'. *Theory and Decision* 38(2): 173-193.

ANDERLINI L. (1990): 'Some Notes on Church's Thesis and The Theory of Games'. *Theory and Decision* 29: 19-52.

ARMSTRONG D.M. (1973): *Belief, Truth and Knowledge*. Cambridge UP.

AUMANN R. (1974): 'Subjectivity and Correlation in Randomized Strategies'. *Journal of Mathematical Economics* 1: 67-96.

AUMANN R. (1976): 'Agreeing to Disagree'. *Annals of Statistics* 4: 1236-1239.

AUMANN R. (1987): 'Correlated Equilibrium as an Expression of Bayesian Rationality'. *Econometrica* 55: 1-18.

AXELROD R. (1981): 'The Emergence of Cooperation among Egoists'. *The American Political Science Review* 75: 306-317.

BACHARACH M. (1992): 'Backward Induction and Beliefs about Oneself'. *Synthese* 91: 247-284.

BALZER W. & TUOMELA R. (1997): 'A Fixed Point Approach to Collective Attitudes'. In *Contemporary Action Theory. Vol. II*. (Eds.) Holmström-Hintikka and Tuomela. Kluwer Academic Publishers

BERNHEIM, D: (1984): 'Rationalizable Strategic Behavior'. *Econometrica* 52: 1007-1028.

BICCHIERI C. (1989): 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge'. *Erkenntnis* 30: 69-85.

BICCHIERI C. (1990): 'Norms of Cooperation'. *Ethics* 100: 838-861.

BICCHIERI C. (1993): *Rationality and Coordination*. Cambridge UP.

BICCHIERI C. & SCHULTE O. (1997): 'Common Reasoning About Admissibility'. *Erkenntnis* 45: 299-325.

BICCHIERI C. & GREEN M. S. (1999): 'Symmetry Arguments for Cooperation in the Prisoner's *The Logic of Strategy*. (Eds.) Bicchieri, Jeffrey & Skyrms. Oxford UP.

BINMORE K. (1987): 'Modeling Rational Players. Part I'. *Economics and Philosophy* 3: 179-214.

BINMORE K. (1988). 'Modeling Rational Players. Part II'. *Economics and Philosophy* 4: 9-55.

BINMORE K. (1993): 'De-Bayesing Game Theory'. *Frontiers of Game Theory*. (Eds.) Binmore K., Kirkman A. & Tani P. The MIT Press.

BLOCK N. (ed.) (1980): *Readings in the Philosophy of Psychology. Vol.1*. Methuen.

BRANDENBURGER A. & DEKEL E. (1987): 'Rationalizability and Correlated Equilibria'. *Econometrica*, Vol. 55, No. 6: 1391-1402.

BRATMAN M. (1987): *Intention, Plans, Practical Reason*. Harvard UP.

BRATMAN M. (1992): 'Shared Cooperative Activity'. *The Philosophical Review* 101: 327-341.

BURGE T. (1975): 'On Knowledge and Convention'. *Philosophical Review* 84: 249-255.

CURRIE G. (1986): 'Review on Ruben's 'The Metaphysics of the Social World. *The British Journal for the Philosophy of Science*: 127-132.

van DAMME E. (1989): 'Stable Equilibria and Forward Induction'. *Journal of Economic Theory* 48: 476-496.

DE FINETTI B. (1975): *Theory of Probability*. Wiley.

DUFWENBERG M. & LINDÉN J. (1996): 'Inconsistencies in Extensive Games - Common Knowledge Is Not the Issue'. *Erkenntnis* 45: 103-114.

ELSTER J. (1983): *Explaining Technical Change*. Cambridge UP.

ELSTER J. (1985): 'Rationality, Morality, and Collective Action'. *Ethics* 96: 136-155.

GALAVOTTI M.C. (1997): 'Probabilism and Beyond'. *Erkenntnis* 45: 253-265.
GAUTHIER D. (1986): *Morals by Agreement*. Oxford UP.
GILBERT M. (1990): 'Rationality, Coordination, and Convention'. *Synthese* 84: 1-21.
GRICE P. (1989): *Studies in the Ways of Words*. Harvard UP.
GUTTENPLAN S. (ed.) (1995): *A Companion to the Philosophy of Mind*. Blackwell.
HAMPTON J. (1987). ' Free-Rider Problems in the Production of Collective Goods'. *Economics and Philosophy* 3: 245-273.
HAMPTON J. (1994): 'The Failure of Expected-Utility Theory as a Theory of Reason. *Economics and Philosophy* 10: 195-242.
HARDIN R. (1971): 'Collective Action as an Agreeable n-Prisoners' Dilemma'. *Behavioral Science*, Vol. 16: 472-481.
HARDIN R. (1982): *Collective Action*. The Johns Hopkins Press for the Resources for the Future.
HARSANYI J. (1982): 'Games with Incomplete Information Played by "Bayesian" Players, I-III'. *Papers in Game Theory*. Harsanyi, p. 115-70. D. Reidel, Dordrecht, Holland.
HARSANYI J. (1982): 'A New General Solution Concept for both Cooperative and Noncooperative Games'. *Papers in Game Theory*. Harsanyi, p. 211-31. D. Reidel, Dordrecht, Holland.
HARSANYI J. (1982): *Papers in Game Theory*. D. Reidel, Dordrecht, Holland.
HARSANYI J. & SELTEN R. (1988): *A General Theory of Equilibrium Selection in Games*. The MIT Press.
HART H.L.A. (1961): *The Concept of Law*. Oxford University Press.
HINTIKKA J. (1970): 'Knowledge, Belief, and Logical Consequence'. *Ajatus* 32: 32-47.
HOLLIS M. (1994): *The Philosophy of Social Science - An Introduction*. Cambridge UP.
HOBBES T. (1962): *Leviathan* (1651). Ed. J. Plamenatz. London, Fontana.
HUME D. (1978): *A Treatise of Human Nature* (1739). Ed. L.A. Selby-Bigge. Clarendon Press.
JACKSON F. (1987): *Conditionals*. Blackwell.
JACKSON F. & PETTIT P. (1993): 'Some Content is Narrow'. *Mental Causation*. (Eds. Heil J. & Mele A.). Clarendon Press.
KANDORI M. (1992): 'Social Norms and Community Enforcement'. *Review of Economic Studies* 59: 63-80.
KAPLAN M. (1996): *Decision Theory as Philosophy*. Cambridge UP.
KEYNES J.M. (1921): *A Treatise on Probability*. MacMillan, London.
KREPS D., MILGORM P., ROBERTS J. & WILSON R. (1982): 'Rational Cooperation in the Repeated Prisoner's Dilemma'. *Journal of Economic Theory* 27: 245-52.
LAGERSPETZ E. (1995): *The Opposite Mirrors. An Essay on the Conventionalist Theory of Institutions.* Kluwer Academic Publishers.
LEWIS D. (1969): *Convention - A Philosophical Study*. Harvard UP.
LEWIS D. (1976): *Counterfactuals*. Basil Blackwell.
LEWIS D. (1983): *Philosophical Papers, Vol. 1*. Oxford UP.
LUCE R.D. & RAIFFA H. (1957): *Games and Decisions*. John Wiley & Sons, Inc.
MARMOR A. (1996): 'On Convention'. *Synthese* 107: 349-371.
MELE A. (1992): *Springs of Action: understanding* intentional behavior. Oxford UP.
MILLER S: (1990): 'Rationalizing Conventions'. *Synthese* 84: 23-41.
NELSON R.J. (1982): *The Logic of Mind*. D. Reidel Publishing Company.
Von NEUMANN J. & MORGENSTERN O. (1944): *Theory of Games and Economic Behavior*. Princeton UP.
NIDA-RÜMELIN J. (1997): *Economic Rationality and Practical Reason*. Kluwer Academic Publishers.
NIETZSCHE F. (1896): *Jenseits von Gut und Böse. Vorspiel einer Philosophie der Zukunft*. von C. G. Naumann. Leipzig.

Political Studies 30: 350-370.
TUOMELA R: (1988): ‘Free-Riding and the Prisoner’s Dilemma’. *The Journal of Philosophy* 85: 421-427.
TUOMELA R. (1992): 'On the Structural Aspects of Collective Action and Free Riding'. *Theory and Decision* 32: 165-202.
TUOMELA R. (1995): *The Importance of Us: a* philosophical study of basic social notions. Stanford UP.
ULLMANN-MARGALIT, E. (1977): *The Emergence of Norms*. Oxford UP.
VANDERSCHRAAF, Peter (1995b): 'Endogenous Correlated Equilibria in Noncooperative Games'. *Theory and Decision* 38: 61-84.
VANDERSCHRAAF, Peter (1995a): 'Convention as an Correlated Equilibrium'. *Erkenntnis* 42: 65-87.
VROMEN J.J. (1995): *Economic Evolution*. Routledge.
VROMEN J.J. (1998): 'If Conventions are Solutions, What Are the Problems?' A Manuscript.

WILLIAMSON T. (1992): 'Inexact Knowledge'. *Mind* 101: 217-242.
YOUNG H.P. (1993): 'The Evolution of Conventions'. *Econometrica* 61: 57-84.