# Rules, Choice, and Economic Organisations:
## An Inquiry into Procedural and Consequential Considerations of Rules

Jukka Kaisla
Department of Industrial Economics and Strategy
Copenhagen Business School

Ph.D. (Econ) Thesis
November 2001

# Contents

## Summary

This study is about modelling rule following, rule making, and rule change in economics in general and with regard to economic organisations in particular. It is maintained that rational choice theory is silent about rule following behaviour even though such behavioural patterns represent a ubiquitous part of human action. An approach called *rule-individualism* is taken to represent a good attempt in providing a behavioural foundation alternative to rational choice theory. The present study will claim, however, that rule-individualism does not succeed in its attempt to replace the *logic of choice* by a genuine *theory of behaviour*. It will be suggested that a behavioural theory that is based on consequential considerations only borrows (implicitly or explicitly) behavioural rules from the maximising framework of rational choice theory. This study will suggest an alternative approach to modelling rule following behaviour that is not reducible to the logic of choice.

The suggested model of rule following behaviour, which is not only based on *consequential* reasoning but also on *procedural* considerations, bears some important methodological implications. It is maintained that a behavioural theory that is limited within a strict interpretation of *methodological individualism* does not provide a realistic picture of reality where rules are not only the outcomes of (inter)action but they also condition behaviour. Procedural aspects of rules direct attention to the behavioural recommendations, cues, and constraints that rules provide. Even though this is a realistic picture of human behaviour, economic literature has so far been largely silent about non-consequential aspects of choice behaviour.

After analysing rules at the individual level, the study turns to look at how rules emerge and change through social processes. The approach is individualistic in the sense that by efficiency it is always referred to the preferences of the people involved. This perspective rejects supraindividualistic criteria of goodness. *Constitutional economics* provides a systematic and consistently individualistic approach to collective action. Its applicability to studying organisational rule making and choice behaviour will be analysed. The present study suggests that while constitutional economics can indeed contribute to the study of economic organisations, it may gain in consistency by introducing *conventions*. The rationale for this argument lies in the strong methodological inclination of constitutional economics. The present study claims that inferring *voluntary agreement* as the ultimate criterion of goodness in constitutional economics from the concept of *market exchange* is ill-suited to a method that is based on strong logical reasoning. The study maintains that voluntariness in social contract relates more to conventions than to the idea of market exchange. The logic for this argument will be provided.

The way *Prisoners' Dilemmas* are interpreted varies in economics. Game theory provides the basic logic, but remains silent about how the values of payoffs become defined. Procedural reasoning, as examined in the study,

alleviates the assumption of overtly consequentially maximising players. This is in line with other studies referred to here which show that cooperation is the default behavioural pattern and that PDs are generally less pervasive in economic organisations than what is assumed. The study will maintain that in economic organisations PD games become transformed into coordination ones and that the members have interests in general conformity to PD rules as well. Since an organisation's constitution comprises not only coordination but also PD rules, organisation designers should pay especial attention to the constitutional order of the organisation when modifications are made. Organisational conventions carry information about how fairness is interpreted in particular contexts.

# Introduction

> It is not the purposive but the rule-governed aspect of individual actions which integrates them into the order on which civilisation rests (Hayek, 1978, 85).

A central motivation for this study is based on the observation that rule-following behaviour provides a major challenge to economics. The concept of rational action seems to be at odds with a behavioural pattern in which the actor does not take into account all the relevant factors in a situation and pursues her way seemingly motivated by other than consequential issues in her mind.

A central problem with orthodox economics is that rule following as a general and prevalent behavioural mode cannot be arrived at from the maximising framework where the individual is constantly engaged in comparative calculation among choice options. The primary aim of this study is to examine how to distinguish rule following from situational judgement in a way that would not be reduced to the logic of choice. The model of a rule following individual is seen here to be incomplete without reference to the structure of social rules. The discussion will thus lead to the arenas of collective rule making and spontaneous rule change. Regarding collective rule making, the study will examine the applicability of constitutional economics in analysing economic organisations. The basic thrust of this study is individualistic, although a strict interpretation of methodological individualism is seen as being problematic. Constitutional economics is thoroughly individualistic providing a set of principles whose logical consistency and applicability will be studied in this study.

The present study will maintain that the constitutional perspective is applicable in the study of economic organisations and firms for both external and internal reasons; the external reason being the introduction of the normative content of rules into institutional economics, while the internal reason refers to a better applicability of constitutional principles in smaller groups, such as firms (compared to countries). However, the present study will maintain that neither rule following nor constitutional economics do well without reference to conventions. Conventions provide a central non-individualistic flavour to this study because even though their emergence and change can be explained individualistically, they also condition behaviour. The complex interplay between the individual, conventions and collective decision making will be examined throughout this study.

A perspective called *rule-individualism* aims at providing a rationale for the individual's behaviour that seems different from case-by-case calculation (Vanberg 1994). The rule-individualistic approach maintains that a central rationale for rule following is the cognitive limitation of the human mind. People cannot pursue case-by-case calculation simply because they lack the

necessary data and processing capacity in any given situation. Thus, they resort to rule following. An implication of this is that *all* action is essentially based on rule following at some cognitive level. But a question then arises: what distinguishes rule following as an observable behavioural regularity from situational judgement if all action is based on rule following?

Assume that cognitive limitations *do not* provide the rationale for rule following. If this is the case, another problem arises: if the individual's behavioural repertoire is not limited to rule following and she has the capacity to assess expected outcomes case by case, why would a rational person resort to rule following in any situation other hand when her weakness of will needs to be mended?

Answering the above question reveals a potential inconsistency of the rule-individualistic explanation as well. The rule-individualistic approach maintains that cognitive limitations provide the rationale for rule following, but another rationale is provided as well: rules are followed 'if rule-following can be expected to result in larger overall pay-offs (over a relevant period of time) than case by case adjustment' (Vanberg 1994, 17). By arguing this, the rule-individualistic position not only refutes the cognitive limitations explanation, but also seems to assume a calculative ability that goes beyond the requirement for case-by-case adjustments. The individual needs to be able, not only to assess separate situations, but more importantly, to form expectations based on the comparative payoffs between long-term rule following and case-by-case adjustments that will never materialise.

The present study aims to examine a complementary rationale for rule following. Explanations based solely on consequential reasoning seem to be at odds with the cognitive limitations of the human mind. If rule following requires long-term precommitment, how can we assume the decision maker to be able to assess the long-term consequences of following a particular rule, not to mention those of which are never chosen.

Rationality can be approached from two perspectives. Individuals act either according to expected outcomes, or alternatively they act according to rules (Hayek 1978, 84). The former can be defined as consequential rationality and the latter as procedural rationality (cf. Simon 1976, 1978, 1979; Le Menestrel 1997). The present study will discuss relations between these two perspectives. A central issue that will be emphasised is that the above distinction between the two types of rationality does not provide a complete picture. This is because a choice among rules can be based either on procedural or consequential rationality. This study will emphasise the fact that even with rule following behaviour, *interpretation* plays an important role.

Prisoner's Dilemmas (PDs) represent social situations in which the participants would prefer to cooperate but remain reluctant to do so because of the fear that others may take advantage of their cooperative behaviour. To be sure, this interpretation is not necessarily shared by theorists who assume that defection is a deterministic response by every rational agent. The discussion of PD rules revolves around the way we picture the players

and their interaction. Findings in experimental and evolutionary economics imply that PDs do not represent so vast and central a problem as is normally assumed in economics (Rabin 1993, Ledyard 1995, Sally 1995, Camerer and Knez 1997, Dosi et al. 1999). This study discusses a conceptual framework that could help us to understand why individuals systematically provide spontaneous resolutions to PDs even when it should be against their immediate self-interest to do so.

David Hume (1969 [1740], 1987) emphasised conventions as guides that help us in finding mutually beneficial solutions to social problems. After his contribution, the interpretation of conventions has changed and today only coordination rules are generally regarded as conventions (since Lewis 1969). This study aims at reviving the type of conventions that are generally not counted as such, namely, PD rules. The rationale for this revival can be found in the procedural justification process that constitutional economics is based on. My aim is to argue that conventions play a central role in facilitating a social contract.

A common train of thought in economic literature maintains that as the individual is best seen as self-interested with guile (Williamson 1985), PDs would remain unsolved without the stabilising effect of enforcement by a third party. And as third-party enforcement is a common mechanism, the self-interest assumption seems correct. This logic can, however, be seen differently as well. The use of third-party enforcement *per se* implies that the individual not only values general conformity to non-conformity, but also wants to tie not only the hands of others but her own as well to ensure general conformity. If individuals were generally self-interested with guile, a third party would not be able to stabilise general conformity because the same guile that is supposed to be directed toward fellow citizens would also be directed to the third party. Thus, nobody would observe the third party to start with. The underlying explanation for stability of PD conventions is not third-party enforcement, but the interests of the members in maintaining conformity. This issue will be analysed further in chapter 5.

### Conventions of terminology in the study

*Consequential interests* explain a type of choice behaviour where the expected consequences of choice options direct the attention of the decision maker. Rational choice theory is consistent with a behavioural model where the individual is solely motivated by the expectations of consequences of choice options. Choosing to follow a rule may also be based on consequential interest if the choice is based on comparative assessment of expected outcomes of alternative modes of behaviour.

*Procedural interests* explain a type of choice behaviour where the expected consistency between a choice situation and the hierarchy of rules direct the attention of the decision maker. The decision maker's attention is

directed to assessing which of the perceived rules best corresponds with the present situation and how to apply a chosen rule in that situation.

*Consequential efficiency* is assessed by the degree to which an outcome of a choice coheres with the expectations, as judged by the relevant participants (cf. 'the single person equilibrium' in Hayek 1948, 36). This perspective is consistent with an *ex post* version of Pareto-efficiency where the observed outcome is assessed good by the participants involved.

*Procedural efficiency* is assessed by the degree to which action corresponds with the rules that are expected to govern the activity in question. Procedural efficiency has some non-individualistic connotations as conventions may provide the benchmark to which a choice of alternative actions is measured. It may deserve mention that to the extent that conventions explain the attainment of a mutual agreement within a group, a social contract can never be strictly individualistic either.

*Rules* are interpreted in the present study as behavioural patterns or regularities of conduct (Hayek 1967, 66) of a person or of a group of people. Rules also refer guides for behaviour or constraints in particular situations.

*Institutions* have the same connotation as rules, but refer to the social level. Institutions often also refer to a structure of rules.

*Social contract* refers to a unanimous agreement within a group of people who are engaged in an organised, collective endeavour. A social contract defines the terms of their participation regarding three main issues: it specifies the resources that the members are to contribute to the common use; it specifies the participants' decision-making rights over the use of the common resources, and how the collective outcome is to be shared among the participants (Vanberg 1994, 220).

*Conventions* bear related but slightly dissimilar interpretations in the discussions throughout the study. A central, invariant aspect of conventions is that they emerge and are maintained by mutual, shared expectations of behavioural patterns among the participants. For some, conventions are limited to coordination rules, whereas for some others, they include Prisoner's Dilemma rules as well. Furthermore, the extent to which conventions are seen as emerging spontaneously is open to interpretation. Generally speaking, conventions emerge and change without intentional design by some agency. Conventions relate to social contract in that they affect the rules by which a division of decision-making rights and an allocation of a collective outcome are arrived at.

*Economic organisations* are seen here as corporate actors that are defined by the following features: the members combine certain resources that are used jointly subject to certain procedural rules. These procedural rules provide the common denominator that coordinates organisational interaction among the members and can be seen as a constitution (cf. Coleman 1974, 1986, 1990).

*Agreement* comprises expectations of the future performance of promises. Agreement is seen here as always being directed towards the

future. If expectations of future performance were missing, agreement would lose its empirical content. This distinguishes agreement from exchange which does not necessarily have any content towards the future.

**The kinds of efficiency that will be considered**

Efficiency in economic organisations generally refers to a consequential interpretation of the concept, that is, that efficiency is assessed by the *outcome* of any given action. There are numerous ways to describe organisational efficiency (profits, growth, etc.), but the common denominator is the emphasis on the outcomes that an economic organisation produces.

An economic organisation is essentially a collective endeavour. From the subjectivist perspective a consequential interpretation of efficiency does not alone guarantee that the members perceive an outcome as mutually desirable to all of them. This is because their interests vary temporally and are not the same across individuals.

Economic organisations do not produce outcomes in an institutional vacuum. The members of the organisation adhere to certain sets of rules that contribute to the organisational processes by which outcomes are brought about. Also, economic organisations operate in a socio-economic environment whose institutional structure conditions activities. Thus rules and institutions become central to the study of economic organisations.

The fact that rules are prior to action because action takes place within a framework of rules poses important questions regarding the connection between rules and their contribution to efficiency in economic organisations. An action can be seen as being efficient insofar as the actor is able to carry out her plan as anticipated (Hayek 1948, 36). This type of efficiency consideration is directed to the consequences of actions. In organisational literature, efficiency may be related to the degree to which organisation members conform to the organisational goals (cf. Merton 1940, Selznick 1948). Insofar as organizational rules transform the goals into behavioural guidelines, the degree of conformity to the rules becomes a central reference point for efficiency.

The contractarian position is claimed to systematically be able to extend the individualistic perspective of classical liberalism into the realm of collective choice (Vanberg 1994, 204). The individualistic position maintains that voluntary exchange indicates agreement among the parties, and that such voluntary agreement is the ultimate criterion on which an exchange can be judged to be efficient (Buchanan 1977, 128). In direct analogy, the contractarian individualistic position maintains that a collective choice can only be judged efficient if it is based on voluntary agreement by all parties involved (Vanberg 1994, 204).

The present study will suggest that even though the efficiency of a market exchange may be assessed by its voluntariness, the use of the analogy

in the contractarian position is not necessarily unproblematic. The efficiency consideration of a voluntary market exchange is limited in that the institutions that constrain such an exchange are taken as given. With social contract things are not as simple because an agreement needs to be assessed against prior rules that define the boundary between voluntariness and coercion. The present study will maintain that the contractarian claim for priority of social contract over conventions makes the position logically inconsistent. A solution is suggested by introducing conventions that provide the required distinction between voluntariness and coercion to facilitate a social contract.

### Constitutional and compliance interests

Vanberg has made an important distinction between the individual's *constitutional* and *compliance* interests in agreeing upon and conforming to constitutional rules (1994, 21-3). The individual's constitutional interests determine what she would prefer if she were to participate in choosing a constitution, whereas her compliance interests are her preferences over alternative choice options, given the constitutional constraints. These interests need not cohere, as the individual may well prefer a constitution based on private property and yet she may prefer to violate the property rights of others.

### Procedural and consequential interests

The theme of this study refers to a social reality in which both constitutional and compliance interests are not only affected by the individual's *consequential* considerations but also her *procedural* interests over choice options. An actor's consequential interests determine what alternative she would prefer based on the expectations of their outcomes alone. Procedural interests determine what alternative stands out in providing coherence between action to be taken and the set of constraints that the decision maker faces. Procedural interests direct the decision maker's attention to the assessment of which rule to choose from the set of alternatives, and how to interpret its meaning in a particular context.

### Methodological issues

This study discuses the applicability of methodological individualism (MI), normative individualism (NI), and subjectivism in the study of social rules. A strict version of MI is considered inapplicable when analysing rule following from the present perspective.

*MI* maintains that social phenomena at the aggregate level should be explained by reference to the actions and interactions of individual actors

who, separately or jointly, pursue their interests as they see them, based on their own understanding of the world around them (Vanberg 1994, 1).

*NI* is based on a normative presumption that the values and interests of the individuals involved provide the relevant criterion against which the goodness of alternative choice options are to be measured (Vanberg 1994, 1).

*Subjectivism* refers here to the independence and subjectivity of preferences of the decision makers. The creative and imaginative capacity of the individual permits knowledge changing endogenously leading to dissimilar knowledge among the individuals. In rational choice theory a subjective element of a choice are the preferences that, together with side constraints, determine the maximising alternative. A more dynamic approach to subjectivism maintains that the individual's choice is not conditioned by what are considered objective circumstances (preferences and constraints). Rather, a choice is originative in the sense that it presupposes radical creativity and independence from the circumstances that exist prior to a choice (Kirzner 1992, 122-3).

In sociology as well as in new institutional economics, the interdependence of the institutional structure and the individual is no news these days. On the one hand, rules and institutions condition choice behaviour, and on the other, rules and institutions themselves are intended and unintended consequences of individual choices. Irrespective of this reasonable picture of the interaction between different levels of social reality, the model of the individual in economics still seems rather disconnected from such an interactive overview. This is to say that even though some interplay may be assumed, the starting point is an autonomous agent who is best pictured as being self-interested with guile, as in Williamson's work. An attempt to bring the impact of social rules upon the individual may of course be dismissed by reference to parsimony.

This may create problems when social rules are analysed by MI. If the picture of the central agent is distant from reality to start with, the rules that are assumed to arise by the interaction of such agents may not represent reality. PD rules are an example of this issue. There is systematic evidence that people do not behave the way orthodox economics assumes them to in PD situations (cf., Rabin 1993, Ledyard 1995, Sally 1995, Camerer and Knez 1997, Dosi et al. 1999). The reason why this body of evidence is increasing is not because people have only recently started behaving cooperatively, but because only recently have theorists started to empirically examine whether or not the assumption of self-interest with guile is reasonable.

Thus, if the parsimonious version of MI provides a biased picture of reality, a theorist needs to choose which one she values more: a realistic explanation or the Occam's Razor principle. The present study is interested in realistic explanations and if the model of the individual needs to be 'complemented' by additional assumptions, then that must be part of the task. The notions of additional or complementary assumptions bear negative

impact by default. One may, however, approach this issue from a different angle by questioning the meaningfulness of stripping the model of the human being to the extent that it fails to provide a picture of appropriate behaviour.

The present study is based on MI in the sense that it is seen as providing the primary direction of explanation, but not the full picture of causality. MI suggests that social phenomena at the aggregate level should be explained by reference to the actions and interactions of individual actors who, separately or jointly, pursue their interests as they see them, based on their own understanding of the world around them (Vanberg 1994, 1). This interpretation is silent about how the interaction between different levels of social reality, between the individual and the institutional structure, should be taken into account.

The present inquiry into the world of rules is based on an epistemological assumption that a rational, consequentially motivated choice among rules is more difficult to arrive at than a similarly motivated choice among available alternatives within an established institutional framework. If we have epistemic problems in assessing the comparative goodness of various choice options within a relatively stable framework of rules, then we certainly have even greater difficulties in assessing the consequential goodness of alternative rules that are logically prior to any choice within a framework that is yet to be established. Thus, it is reasonable to assume that the participants turn towards procedural considerations in their search for proper rules.

The present approach accepts the contractarian normative position which holds that the rules that emerge from an acceptable process, that is, from a process based on mutual agreement, are, by inference, acceptable (Buchanan 1977, 293). However, it is also maintained here that the limits that people have in assessing long-term consequential efficiency of rules directs their attention toward procedural considerations. This is to say that while the contractarian principle provides the normative criterion for the procedural assessment, the present study aims to explain why and how people come to recognise such a criterion.

**Order of treatment**

The study will proceed as follows: chapter 2 will examine the model of the economic agent from the rule following perspective. A general observation suggests that the rationally maximising representation of the individual is strongly present in economics. If we allow the individual to follow rules, then at least her choice of doing so must be based on some, albeit limited, calculation of expected outcomes (the consequential perspective). It is suggested that rule following as a behavioural pattern may owe more to non-consequential considerations than to consequential assessment between alternative behavioural modes.

Chapter 3 examines conventions. The purpose of this chapter is to analyse the central principles and dynamics that give rise to conventions and their change. A central issue with coordination rules is how prominence is viewed. The perspective of the present study emphasises the interpretation element in prominence.

Prisoner's Dilemma (PD) rules are also considered conventions, and it is maintained that the stability of PD rules is established by mutual expectations, just like in social contract. If devices are established to ensure stability, such devices need to satisfy the mutual benefit argument. Government can thus emerge spontaneously and can be viewed as part of the spontaneous stabilising mechanisms of PD rules.

There is good reason to assume that a transformation from PD into coordination games is an important part of social interaction. Camerer and Knez (1997) provide some interesting insight into this issue. In this relation, Hume's account of conventions will be discussed. Gauthier's (1998) analysis of Hume's approach shows close affinity between social contract and unstable PD conventions.

My aim in this chapter is to argue that the underlying explanation for the resolution of the instability problems of PD rules is not a social contract. A social contract is an outcome, an end result, of a process by which the instability of PD is resolved. Such a process is essentially about the emergence of a convention.

Chapter 4 examines the principles by which purposefully designed rules are established from the normative individualist perspective. This chapter examines further the procedural efficiency criterion applied in the study. The contractarian tradition, based on methodological and normative individualism, analyses the processes by which a collective endeavour can be assessed as efficient. The present examination discusses the limits of individualistic efficiency criteria. It is argued that voluntary exchange as the ultimate source of goodness in collective choice remains incomplete.

The constitutional approach to the procedural criterion of goodness is justified up to the point at which I am willing to sacrifice the strict version of MI for an explanation that I feel can contribute to our understanding of the individual's choice behaviour. A voluntary exchange is impossible without shared rules that demarcate between voluntariness and coercion. Although exchange of *commitment* is necessary for such rules to become stabilised, it is not exchange *per se* that explains exchange. What facilitates exchange is a process by which the demarcation between voluntariness and coercion becomes derived from the shared sense of fairness over an extensive period of time. That process is essentially the process by which conventions emerge and change. Thus a realistic model of rule making cannot be entirely individualistic.

It will also be argued that efficiency considerations, based on normative individualism, are limited by the normative impact of rules that are conformed to in a community. Opportunity costs are considered a

potential candidate in breaking with the normative content of rules. The present study is, however, unable to find a satisfactory way to view opportunity costs as positive entities, disconnected from the rules by which they are created and changed.

The chapter will also discuss procedural and consequential issues regarding the efficiency of rules. The distinction between the procedural and the consequential approach to efficiency is considered beneficial because it clarifies a central aspect of efficiency: regardless of which perspective (or which combination between these perspectives) one is to choose, the consideration of efficiency always remains a partly subjective task and thus subject to speculation in a social context.

Chapter 5 analyses the interplay between purposeful action and the nature of outcomes. The chapter draws upon the analyses in the preceding chapters. The distinction between procedural and consequential efficiency will be examined further. Hayek's theory of cultural evolution will be discussed in this relation. Two alternative explanations for the development of appropriate rules are considered: the invisible-hand and the group selection explanations.

After Vanberg's critique of Hayek's position, the chapter will examine possibilities for the re-evaluation of Hayek's approach from a non-consequential perspective. The consequential and procedural elements in social contract and the evolution of rules will be discussed. The common law process provides a field where both consequential and procedural considerations are of import. Finally, the connection between purposeful behaviour and unintended consequences at the aggregate level will be discussed.

Chapter 6 examines how organizational rules are examined in the economic literature of organisations. A central finding is that organisations are primarily examined in ways that address the consequential assessment of efficiency. As Nelson and Winter explain, a routine is a failure if it is unprofitable (1982, 121). The present study suggests that even though the consequential assessment of efficiency is a relevant part of human choice behaviour, it does not alone provide a satisfactory explanation for efficiency in collective endeavour. The procedural assessment of efficiency is an important part of justification, especially in collective activities such as economic organisations. Despite this, the procedural assessment of organisational rules is by and large nonexistent.

The analysis by Cyert and March (1963) of organisational decision-making processes implies that prior commitment rather than marginal return play a central role in the assessment of choice options. Organisational expectations are generally influenced by hope, wishes, internal bargaining needs of subunits, conscious as well as unconscious manipulation of information and expectations, and other influencing behaviour (p. 97). Sober, consequential assessment of choice options may in general be such

an idealised mode of behaviour that it hardly exists in the organisational context.

A central finding of this chapter will be that although organisational rules have, to some extent, been analysed in heterodox economics, the literature is largely silent about issues that are emphasised in constitutional economics. These findings imply that there is at least a good starting position for constitutional economics to add value to our understanding of how rules influence organisational behaviour, and especially to enhance our perception of the principles and processes by which rules change.

Chapter 7 examines the applicability of an extended constitutional perspective to business firms. The firm is constituted by a group of self-interested people cooperating and competing within a set of multi-layered rules. Drawing upon the analyses in chapters 3 to 5, the chapter aims to extend the social contract approach to the economic organisation by introducing conventions. The importance of this contribution lies in the procedural aspect of conventions. Without the extension suggested here the attainment of a social contract would remain unviable.

Chapter 8 aims at illustrating the complexity of the interrelations between intentional and unintentional elements in open source software development. The open source software idea is based on the freedom to use, copy, modify and redistribute software. The term *open source* means that the source code needed to modify software is provided, and that the users/developers have the right not only to use, but also to modify and distribute modified versions. The starting point is that nobody is permitted to pronounce an exclusive property right to open source software. The proprietary model with which the open source model is convenient to be compared, is based on a more conventional idea of copyright. The developer/distributor reserves all rights to copy, modify and distribute while users only have the right to use the software.

The elements of the model are examined through their degree of intentional design vs. unintended impact, and their degree of importance or necessity to the process. The story will start from general conventions of fairness in the software-developing community. These conventions will bring about unintended consequences without which the concept of open source software would arguably not have emerged. The conventions of fairness give rise to specific conventions of property in open source development. Drawing upon these conventions, the central players in open source development designed a social contract to maintain the beneficial pattern of cooperation among developers.

Open source software projects have developed in ways that seem to defy some general assumptions regarding consequential efficiency. Open participation has produced technologically high-quality products. The chapter analyses the constitutional dynamics of open source software development. The lack of conventional organisational structure emphasises the working properties of social contracts and conventions. The aim of the

chapter is to illustrate some central dynamics in the interplay between evolution and design.

Finally, chapter 9 will discuss some conclusions based on this study. It is maintained that when modelling rule following as well as the economic agent, considerations of procedural aspects in choice situations may help to form a more realistic picture of rational choice making. A strict interpretation of methodological individualism does not seem to permit an explanation where conventions and other social institutions partly condition the individual's choice behaviour. Procedural aspects in choice behaviour also permit dynamics of interaction among the participants where PDs transform into coordination problems in economic organisations. Viewing organisational social contracts as exchange of commitments based on consequential reasoning alone is found unsatisfactory. Introducing conventions into constitutional economics is maintained to enhance our understanding of the interplay between intentional and unintentional elements in institutional change.

# Chapter 2

# The Rule-guided Individual

# 1 Introduction

The purpose of this chapter is to analyse rule following at the individual level. Institutions can be conceptualised as sets of interconnected and mutually-stabilising behavioural patterns, and can be seen to be constituted by individuals' routine practices (cf. Vanberg 1994, 7). My aim here is to discuss representations of different approaches to rule following and to suggest that a potentially important approach has been left unnoticed.

The theme of this chapter deals with the distinction between procedural and consequential interests of the individual in following rules and choosing among them. Rational choice theory provides a universal explanation for every type of behaviour by demonstrating that individuals prefer better for worse. The present study takes this axiom as its point of departure and relates it to another, perhaps less recognised, methodology of the Austrian School, namely, *praxeology*. Through comparing the central principles of these related approaches I hope to be able to show their applicability as well as their limitations in providing explanations for behavioural regularities.

After the discussion on rational choice theory and praxeology I will turn to analyse the *rule-individualism* approach. Advocates of this approach recognise the limitations of rational choice theory and seek an explanation for our choice behaviour at the level of rules, instead of at the level of separate choice situations. It will be suggested that focusing solely on the consequential interests of the individual, rule-individualism does not succeed in the effort to 'replace the *logic of choice* by a genuine *theory of behaviour*' (Vanberg 1994, 7, emphasis in original), leaving the effort, therefore, incomplete.

In order to be able to provide an explanation for rule following in a way that extends beyond the pure logic of choice, it may be helpful to analyse the working properties of the mind (Hayek 1952) and the individual's psychological dispositions as well (Vihanto 1998). Behavioural regularities exist that represent a variety of behaviour sometimes even contradicting the rationality assumption.

Hayek's theory of mind introduces a serious problem to the analysis of human behaviour. The central conclusion of the theory is that every type of action is essentially based on our categorising ability. Thus, all action is rule-based. But my intention here is not to simply change the label from rational choice to rule-based choice and pursue the same old picture of the rationally maximising agent. I will suggest that there is more to rule following than meets the rational agent's eye. Hayek's theory of mind does not differentiate between rule following and case-by-case adjustment as *observable* behavioural alternatives. What it does is it gives reason to view even case-by-case calculation being based on categorising.

The distinction between procedural and consequential interests, introduced in this study, may help to understand some important differences between rule following and situational judgement. A behavioural pattern turns out to be of a different type when the individual is motivated by her interest in finding a proper rule and a proper interpretation to it, than when her interest is focused on expected consequences of available choice options. A central issue is then that the individual's cognitive capacity is not the decisive factor differentiating between rule following and situational judgement.

# 2 Rational Choice Theory

Rational choice theory (RTC) recommends forming explanations of social phenomena beginning with the individual's choice behaviour which is viewed as being rational. Rationality manifests itself in the purposeful action of the individual. The unit of analysis is a choice made by an individual. The theory suggests that an individual will always choose an alternative that maximises her utility (or minimises her disutility) in any given situation (cf. Coleman 1990, 13-9). What these choices are about does not concern RTC as it is essentially about how well means are applied in the pursuit of ends, rather than defining the aims of individuals (Elster 1986, 1). An action is rational if the individual has reason to believe that her chosen course of action is the best means to attain whatever she is aiming at. This version of RTC is essentially *subjectivist*.

A choice situation in the rational choice theory can be decomposed into three separate elements: (1) a feasible set of alternatives, (2) a causal structure that connects the feasible alternatives to their outcomes, and (3) a subjective ranking of these feasible alternatives (Elster 1986, 4). A feasible set can be interpreted as comprising objectively existing alternatives that the individual may or may not perceive, or it can be interpreted as comprising only those alternatives that the individual actually perceives. The former, an objectivist version of RTC, does not necessarily provide behavioural recommendation since the individual's behaviour is not affected by what she does not perceive. What the objectivist interpretation claims, however, is that we can make predictions by examining the constraints of a choice situation. This approach is based on assumptions that individuals' preferences are stable both intraindividually and across individuals, thus their theories about the consequences are identical (Vanberg 1994, 26).

A causal structure that links the choice-options to their respective consequences is viewed in the subjectivist version as a set of *theories* about what will happen if the agent chooses this alternative or another (for a Popperian account, see Harper 1996). The causal structure provides the chooser with expectations about alternative consequences, expectations that can be based on false theories.

The objectivist version of rational choice theory appears unsatisfactory as its epistemic requirements conflict blatantly with reality. The subjectivist version on the other hand tends to make the rational choice theory irrefutable. Gary Becker's comment illustrates this:

> According to the economic approach, a person decides to marry when the utility expected from marriage exceeds that from remaining single or from additional search for a more suitable mate (Becker 1976, 10).

This type of explanation provides the logic of choice in the same way as *praxeology*, the theory of human action, does (Mises 1949). In the next section, I will shortly examine the central ideas of praxeology to illustrate the similarities and limitations of these two approaches.

### *1.1 Praxeology, a universal theory of human action*

The praxeological theory of human action is a system of propositions that is based on the methodology of *apriorism*. The key to understanding economic regularities is to regard them as consequences of purposeful actions of individuals. Empirical findings are not needed as theorems can be deduced from the 'knowledge of the essence of human action' (Mises 1966, 64). 'The only way to a cognition of these theorems is logical analysis of our inherent knowledge of the category of action' (ibid.). Just like in rational choice theory, the ends of individuals are taken as given, as beyond the scope of inquiry: 'Praxeology is indifferent to the ultimate goals of actions. Its findings are valid for all kinds of action irrespective of the ends aimed at. It is a science of means, not of ends' (ibid., 15).

The core axiom of praxeology is that 'human action is purposeful behavour' (Mises 1966, 11). Individuals act purposefully in the sense that they can expect that their choices and actions contain the property of affecting the outcomes. This does not, however, mean that the individual should know the precise nature of the outcomes before they unfold. Purposefulness is important for human action because if the individual should believe that there is no connection between her choices and the outcomes, it would become disadvantageous for her to act at all. Every action would *only* inflict cost as the benefit of the outcomes would accrue irrespective of whether or not she acted at all. What makes purposefulness a logical axiom is that any attempt to deny it would be self-contradicting because a counterargument requires purposeful action in itself.

Mises defines the logic of choice much in the same way that can be found in the subjectivist version of RCT:

> Acting man is eager to substitute a more satisfactory state of affairs for a less satisfactory. His mind imagines conditions which suit him better, and his action aims at bringing about this desired state. … There is no standard of greater or lesser satisfaction other than individual judgements of value, different for various people and for the same people at various times (Mises 1966, 13-4).

Mises, however, has his distinct view about rationality: 'Human action is necessarily always rational. The term "rational action" is therefore pleonastic and must be rejected as such' (ibid., 19). There appears to be an unavoidable trade-off between a theory's universality and its explanatory power (see, e.g., Buchanan 1979). Praxeology and rational choice theory are

universal in providing an explanation for every type of choice behaviour. But that universality does not have much predictive power unless preferences and constraints are defined. Since the principle of preferring better to worse does not contain much information, the preferences and constraints carry most of the predictive load.

## *2.1 Subjectivist and objectivist defences of rational choice theory*

### A subjectivist defence

Rational choice theory can be criticised because individuals appear to behave in ways that do not correspond with the assumption of strict self-interest. Psychological and social elements, such as power, trust, altruism, morality, conventions and culture arguably influence choice behaviour but cannot be analysed by RCT. Rational choice theorists may defend themselves against this type of criticism by noting that individuals act rationally *within constraints* and that RCT does not even attempt to define the institutional environment where a particular action takes place. They may suggest, for instance, that we do not have to abandon the assumption of self-interest to explain why individuals may behave altruistically. RCT can be defended by resorting to the universal law of rational behaviour which demonstrates that whatever alternative an individual chooses, she does it because she expects it to maximise her utility, and there is no reason to assume that her utility could not include feelings and beliefs that are not obvious to anybody else (such as self-sacrifice and altruism).

### An objectivist defence

Another type of RCT defence is suggested by Becker (1976). His aim is to show that the objectivist approach to rationality is in fact plausible. His point is to argue that the subjective and variable preferences of individuals are actually based on objective and universal preferences that are stable and identical across individuals (ibid., 5). The same type of idea can be found earlier in Menger's text:

> The maintenance of life depends neither on having a comfortable bed nor on having a chessboard, but the use of these goods contribute, and certainly in very different degrees, to the increase of our well-being. Hence there can also be no doubt that, when men have a choice between doing without a comfortable bed or doing without a chessboard, they will forego the latter more readily than the former (Menger 1950 [1871], 123)

The problem with the objectivist defence is that we do not know individuals' separate, *subjective* theories about the world and how they

connect the alleged fundamental and objective underlying preferences, say, health to the subjective situational preferences, e.g., eating vegetables instead of taking up physical exercise. Becker's objectivist theory seems to replicate the subjective version as well:

> In the standard theory all consumers behave similarly in the sense that they all maximize the same thing – utility or satisfaction. It is only a further extension then to argue that they all derive that utility from the same "basic pleasures" or preference function, and differ only in their ability to produce these "pleasures" (Becker 1976, 145).

This is precisely what is demonstrated in the pure logic of choice approach. Vanberg's (1994, 28) claim that Becker's theory fails in its attempt to establish an objectivist version of rational choice theory seems, therefore, justified.

On the other hand, objectivist aspirations seem not unjustified altogether. It may be that what we consider objective corresponds better with behavioural regularities, with rules and institutions, than with individuals' preferences. What facilitates decision-making is the fact that most rules and institutions that guided our actions yesterday are there also today. We do not have to rediscover the world anew at every instant. This is to say that our relations with the social environment have a *parametric*[1] feature as well. Individual choices have an effect on the market process, but because of the large number of actions, and especially because many of these actions are guided by rules, our relation with the environment can be viewed as being partly parametric. Hayek presents the connection between the individual and the structure of rules as follows:

> The mind is embedded in a traditional impersonal structure of learnt rules, and its capacity to order experience is an acquired replica of cultural patterns which every individual mind finds given (Hayek 1979, 157).

Witt (1991) is in the same line of reasoning when he writes that the variety of individual preferences can be explained as a result of an on-going learning process, the current environmental conditions determining to what extent individually chosen actions are rewarded or punished (ibid., 567).

These ideas give individual preferences a different appearance compared to the radical subjectivist approach. People learn to know what

[1] Parametric is here understood as a type of relation between the individual and her environment where the individual does not have to consider the consequences of her choices on the choices of other participants. In contrast, a strategic relation would be one where the individual needs to consider the consequences of her choices on other participants' choices.

they want and this ongoing learning process is conditioned by the parametric environment.

The facts that preferences vary across individuals, that situational constraints differ and that individuals' cognitive capacities and their theories about consequences vary, do not prevent us from trying to make the world intelligible to us and, to some extent, to make predictions about future events. But it can be reasonably argued that rational choice theory cannot provide what is needed in these processes. The above discussion suggests that RCT cannot solve the dilemma of becoming either irrefutable (in its subjectivist version) or unrealistic (in its objectivist interpretation). Individuals' actions are not made intelligible and predictable by deriving them from rational choice theory only. In order to accomplish these goals, we have to look elsewhere in our search for behavioural theories that can explain social regularities beyond the pure logic of choice. I will now turn to examine theories of rule following as alternatives for rational choice theory.

# 3 Rule Following as a Rational Choice

Judging by the contents of the pure logic of choice approaches, it is reasonable to argue that an alternative theory that could explain the rise of behavioural regularities does not have to penetrate very deeply into the psychological dispositions of the human being in order to outweigh the empirical contents of the former. The justification of this claim lies in the very universalism of the pure logic of choice approaches. By trying to explain everything they fail to explain anything.

## *3.1 Rowe's model*

Rowe (1989) argues that what he calls 'act-individualism', the behavioural description of rational choice theory, cannot explain socio-economic regularities, such as rules and institutions. 'If act-individualism were true, then social facts, social institutions, society, could not exist' (p. 4). This is because a self-interested maximiser would be unable to forego an opportunity to defect while others signal willingness to cooperate.

Therefore, we will need an alternative behavioural theory to explain these regularities. For Rowe, the alternative is found in applying rational choice at the level of choices among *rules* of actions, instead of at the level of choices among actions themselves. 'A rule of action is rational if, by following that rule, an agent maximizes his expected utility' (Rowe 1989, 5). A single action cannot be judged rational as such, but only by considering to what extent it corresponds with a rule that is rational to follow. He concludes that 'social institutions are in fact nothing more than agents rationally following rules of action, and being believed by other agents to do so' (ibid.).

Rowe's rationale for rule following is based on an appealing idea for any rational choice theorist: if individuals are rational when buying and selling, then why should they not be rational in other activities, like in choosing whether or not to follow a certain rule? He explains the basic logic of rational choice among rules as follows:

> Whereas act-individualism proposes a one-step test of rationality, the action being evaluated directly in terms of its consequences, rule-individualism proposes a two-step test of rationality, the action being evaluated in terms of the rules to which it conforms, and the rule in turn being evaluated in terms of the consequences of following that rule (Rowe 1989, 23).

Therefore, Rowe concludes that:

> [i]f the value to an agent of violating his rule exceeds the value to him of maintaining his reputation for following it, then he will violate that rule (Rowe 1989, 24)

Vanberg (1994, 31-2) notices that Rowe's rationality assumption is based on a kind of 'second order' case-by-case calculation in the sense that the individual, instead of calculating which choice-option is rational to choose, calculates whether or not following a rule is rational. If violating a rule gives larger expected pay-offs than following that rule, then the individual will defect.

There seem to be some logical problems in Rowe's reasoning as well:

- Postulate 1: an action is rational only in so far as it is part of a rational rule of action – it is neither rational nor irrational in itself (Rowe 1989, 5).
- Postulate 2: a rule of action is rational if, by following that rule, an agent maximizes his expected utility (ibid.).
- Hypothesis: if the value to an agent of violating his rule exceeds the value to him of maintaining his reputation for following it, then he will violate that rule (ibid., 24).

Insofar as postulate 1 holds, any action that is not part of a rational rule would be neither rational nor irrational. Violating a rational rule would then also be neither rational nor irrational. As change in rules often requires a violation of some existing rule, a change in rules would then become neither rational nor irrational. If a change in rules becomes neither rational nor irrational, then rules themselves become neither rational nor irrational.

The foregoing hypothesis seems inconsistent with postulate 1. The individual appears suddenly capable of defining rationality of an action that is not part of a rational rule.

Disregarding logical problems, what Rowe is arguing is that the individual compares general experience of a certain rule with the expectations of a particular rule-violation. If violating the rule appears beneficial, the individual will act accordingly. Although this is intuitively a realistic description of how the individual makes certain choices it fails to prove precisely what Rowe is aiming at, namely the rationale for respecting rules like property rights. On the one hand Rowe argues that observing property rights is rational in the sense described above. On the other hand, violating property rights is equally rational if violation maximises expected utility. To be sure, Rowe's model suffers from the same tautological tendency as rational choice theory.

### *3.2 Vanberg on the rationality of rule following*

Vanberg (1994) has also adopted the term 'rule-individualism' to define the behavioural foundations of the individual. The individual is unable to calculate the best course of action in separate, dissimilar situations and therefore adheres to mental processes which are not analysed in rational choice theory. She can use her past experience and her categorising ability to make conjectures about the consequences of her choice-options. Individuals are ascribed with 'the capability to learn from experience, and to adapt, over time, their repertoire of behavioural rules to relevant aspects of their environment' (ibid., 29).

By definition, the goodness of rules cannot be judged by their performance in a single situation. Rule-following means that the individual gives up the desire to evaluate every choice situation as a separate and that she commits herself to the rule that has worked well in the past. This notion is not necessarily shared by all advocates of rule-individualism, however. As above, e.g., Rowe (1989, 23) interprets that the meta-choice between case-by-case calculation and rule following is a continuous case-by-case assessment process where the individual evaluates the potential outcomes of violating a rule against the past outcomes the rule has brought about.

For Vanberg, the essence of rule following is *not* to calculate in every choice situation, but, to some extent, to remain unresponsive to the changing particularities (Vanberg 1994, 33). To say that an individual *chooses* to follow a rule would, therefore, mean that the individual basically possessed the capacity to evaluate situations case by case, but would voluntarily give up her calculative capacity. As Vanberg puts it, 'she would have to decide, by rational choice, not to be rational' (ibid., 34). Thus, for Vanberg, the individual does not seem to possess a capacity to switch between rule following and case-by-case calculation at will.

On the other hand, Vanberg views the rationale for rule following as being based on 'some comparison among potential alternative general patterns of behaviour' (1994, 17). To adopt a rule is then rational if it is expected to be more advantageous than an alternative strategy:

> We can view an individual's adoption of a behavioural rule as being based on some comparison among potential alternative general patterns of behaviour. To adopt a rule in this sense can be considered 'rational' if it is found to be a more advantageous strategy than potential alternatives, where attempting to maximize on a case by case basis can be viewed as *one* alternative. … In general it can be argued that adopting a rule for how to behave in certain types of situations is rational if rule-following can be expected to result in larger overall pay-offs (over a relevant period of time) than case by case adjustment. (Vanberg 1994, 17)

An important question arises about whether or not the individual is, even in principle, able to recognise that rule following will be on balance advantageous compared to case-by-case judgement. The logic of reasoning that I am interested in here is as follows: if the rationale for rule following is based on our cognitive limitations that preclude case-by-case calculation, then rule following describes choice behaviour in general. If, on the other hand, the individual is in fact able to pursue case-by-case calculation but prefers to follow rules, then cognitive limitations do not provide the rationale for rule following.

To find out whether complying with a particular rule is more advantageous than rule-violation in the long-term may be difficult for the individual to establish. The individual needs to evaluate and compare potential consequences of both rule following and rule-violation in order to know whether or not the former is on balance more advantageous. This then would indicate that the individual's cognitive limitations do not explain rule following. A cognitive capacity is actually required in a special sense to arrive at a rational choice to follow a rule. As the 'very nature of rules implies that their "goodness" can only be judged by their performance over a longer sequence of applications' (Vanberg 1994, 29), it becomes unclear how the individual, even in principle, could know when rule following is on balance more advantageous. This problem arises because if the individual has experience about following a rule, then she by necessity lacks the experience about the innumerable situations where she might have violated the rule. Any suggestion that she might know the latter cases (which have never been disclosed) fails to give a reasonable account of the fact that she does not even know the nature of the non-existent violations, that is, she does not know how precisely she might have chosen to violate the rule nor at what particular instant she might have done so, not to mention the possible consequences of doing so. The expected consequences of rule following can be viewed as being limitedly predictable, but only insofar as experiences from a rule have already been accumulated. But to claim that one can evaluate the consequences of future case-by-case adjustments or of those that might have taken place in the past would require mental capacities that are difficult to establish.

On the other hand, Vanberg does not view rule following as necessarily providing better overall consequences than case-by-case adjustment:

> Following a rule rather than adjusting to the particular circumstances of each individual choice situation may involve a trade-off: the savings in decision making costs may have to be paid for by decreased overall 'quality' of choice-outcomes (Vanberg 1994, 18).

Three different types of rationales for rule following have been considered here: (1) cognitive limitations, (2) overall advantageous consequences of rule following and (3) savings in decision-making costs. The question whether or not the consequences of rule following can be viewed as being, on balance, more advantageous than case-by-case adjustments, according to a chosen criterion of goodness, is not entirely unproblematic. One can resort to a functional claim that if a rule exists, then it must be more desirable (and in that sense more advantageous). But this rationale has the same kind of irrefutability character as is found in rational choice theory. Savings in decision-making costs are an obvious consequence of reduced decision-making. An open question remains, however, about how we can balance these savings with the reduction of the quality of outcomes of non-existent activities. This refers to the fact that if the actor decides to follow a rule, then she foregoes situational judgement and therefore cannot know the quality of outcomes that the numerous separate choice situations would have resulted in if they had been chosen. It should perhaps also be noted that the explanation based on overall advantageousness of rule following and the cost-reduction explanation are potentially conflicting. The cost-reduction explanation suggests that case-by-case adjustment would in fact give more advantageous outcomes than rule following, whereas the advantageousness explanation claims the opposite. Both of these explanations are at odds with the cognitive limitations explanation. If the individual is viewed as being incapable of case-by-case calculation in the first place, then it is difficult to see how case-by-case calculation could be viewed as an available behavioural mode to which rule following should be compared.

Cognitive limitations as a rationale for rule following introduces some interesting questions. A central question for the present study is whether or not cognitive limitations discriminate between rule-following behaviour and situational judgement. This question arises insofar as the mental processes regarding both rule following and situational judgement are based on the same classification mechanism that is analysed by Hayek (1952).

### *3.3 Cognitive capacity and ability to switch between rule following and discretion.*

There seem to be at least two interesting questions open here. The first question concerns our cognitive capacity. Vanberg criticises the view according to which the rationale for rule following could be based on calculation, either at the level of choices among actions or at the level of choices among rules of actions. Both rational choice theory and Rowe's version of rule-individualism are thus unsatisfactory in the light of the classification process advocated by Vanberg. Individuals follow rules precisely because they lack the capacity to evaluate separate situations in their full details. Even in a situation where the individual cannot find enough

familiar elements to associate it with any already existing category, she uses the same experience-based classification process to establish a new tentative category. This is to say that when engaging in situational judgement, the individual is in fact using the same classifying process that rule following is based upon. Thus the classification act *per se* does not differentiate between rule following and situational judgement.

The second question introduces some new features to the above discussion as it asks whether individuals also follow rules in types of situations where their cognitive capacities would not prevent them from situational judgement. If the answer is in the affirmative, then cognitive limitations do not provide full explanation for rule following either.

Vanberg (1994, 33) views that rule following requires, to some extent, unresponsiveness toward the particularities of a situation. Unresponsiveness does not, however, necessitate the individual's incapability of evaluating a situation. Whether or not an individual is capable of situational judgement and whether or not she uses this capability are two different questions. Being able to evaluate a situation but refusing to do is not an available option. 'For a person to deliberately choose to follow a rule would require him/her to give up, by wilful choice, her capacity to calculate' (ibid., 33-4).

The present study takes a different perspective to this issue. The distinction between procedural and consequential interest may help to clarify why a person may completely rationally choose to follow a rule retaining her capacity to calculate. If the individual's interests were assumed to be directed toward consequential assessment only, Vanberg's position would be justified. But, if we permit the individual to have interests directed to the procedural assessment, the picture changes. A central point to my position is that rule following requires *interpretation*. The individual needs to decide which rule to apply in a particular situation, how to interpret its meaning. If that interpretation act is directed towards consequential issues, that is, to figuring out which rule provides the best expected outcome, the individual is not giving up her ability to calculate, but is comparing the expected benefits that alternative rules would bring about. In procedural assessment, the benefits are not derived from expected consequences of alternative rules, but instead, from their expected appropriateness in a given situation.

Thus the individual does not give up her capacity to calculate; rather she is using that capacity motivated by the two types of interests, the procedural and the consequential. This perspective permits the possibility for the individual to switch between rule following and situational calculation in the sense that she is able to switch between procedural and consequential interests. This view fits well into the picture when we consider rule change. Changing rules requires some initial deviation or innovation. If the individual would only consider her procedural interests the whole time, rule change would occur as an unintended consequence due to the uncertainty of interpretation. Mistakes or variance in interpretation would give rise to new behavioural regularities. But insofar as the individual may in

a situation where her action has normally been based on procedural consideration direct her interest toward alternative consequences, she may break the regular pattern and discover an entirely novel behavioural solution.

# 4 Perception as a Process of Classification

Rational choice theory describes the choice process as one that begins at the recognition of available alternatives, continues by the evaluation of their respective expected consequences and ends at choosing the most preferable option. It is, however, unable to explain how individuals acquire and use knowledge to pursue rational choice behaviour in the first place. The previous section examined how rule following may be viewed from the rational choice perspective. The aim of this section is to turn the table around and suggest that rational choice behaviour can reasonably be viewed against the background of the rule-following disposition of the human mind.

In the previous section I hinted at the possibility that rule-following behaviour may not be entirely the function of cognitive limitations of the human mind. In this section, Hayek's theory of mind (1952) will be examined, one of its implications being that *every* type of action, whether rule following or case-by-case adjustment, is based on perception formation through the categorising disposition of the mind. The working properties of the mind can thus be specified as rule following. But if the rule-following disposition of the mind does not discriminate between case-by-case adjustment and rule following at the observable action level, then we need to search for an additional explanation of rule following elsewhere.

If cognitive limitations of the human mind do not explain rule following entirely, then a possibility may be left open for the individual to be able to switch between unresponsive rule following and case-by-case adjustment at will. The individual is perhaps not constantly using her full cognitive capacity and may well follow rules unconsciously or habitually. A problem with this type of interpretation is that rule following as an efficient-response-to-genuine-uncertainty type of explanation becomes speculative. Irrespective of such a hazard, this is precisely what will be considered here. It may well be that the pressure from rational choice theory and the maximisation framework distort our view of rule following. All types of behaviour are supposed to be maximising in one way or another. If the agent follows rules, then it must be because that is the best thing she can do under the constraints of limited reason and genuine uncertainty. Anything else would undermine the status of human rationality.

## *4.1 Interpretation*

The theory of mind developed by Hayek (1952) analyses the foundations of the individual's choice behaviour. The important part of Hayek's theory of mind for the present study is the process of perception formation. It examines processes by which the individual becomes aware of events and things, i.e., how she makes the world intelligible to herself.

Central to Hayek's theory of mind is the notion of *interpretation*. This notion is also important for the present discussion because it functions as a bridge between case-by-case adjustment and rule following. The central message of *The Sensory Order* is that every type of action, including rule following, requires constant interpretation. In the previous section it was discussed that action, in order to qualify as rule following, needs to be unresponsive toward the particularities of the event the agent finds herself in. On the other hand, the agent faces a problem of choosing which rule to follow at particular types of events. She has to interpret the situation even before a rule-following type of behaviour can commence.

A problem with the idea of constant interpretation is that individuals seem to also follow rules which they are not conscious of. Rules do not necessarily exist in articulated forms (Hayek 1973, 43), or even articulable forms (Hayek 1952). In such cases, interpretation becomes rather an innate process of the mind as the individual may remain unaware of any interpretative effort. This feature relates to an interpretation of rules as behavioural patterns or regularities of conduct (Hayek 1967, 66). If rules are viewed as observable recurrent patterns of behaviour, the problem of constant interpretation does not arise. The only thing that counts then is the behaviour itself, not whether it is an outcome of unresponsiveness to the particularities of events or of some interpretive effort, or any combination of these two.

By interpreting rule following as categorising and rules as outcomes of the categorising act, that is, as recurrent patterns of behaviour, Hayek provides a consistent demarcation between cause and effect. Rules are the result of rule following. Later on in this chapter it will be suggested that the requirement of constant interpretation may need to be relaxed to encompass a certain type of rule following.

### *4.2 Pattern perception*

Hayek's theory suggests that what we can perceive are the *recurring patterns* of separate situations (1967, 23). What our mind is trying to figure out when we are faced with a new situation are elements that show some resemblance to those that we have experience of. We are trying to find possible connections between the elements of the situation we find ourselves in and the categories we have accumulated through experience.

The human mind is, however, limited in the sense that we cannot go through the innumerable particularities of a new situation and compare them separately as an automaton with our cumulated experience to find common elements. The human mind is not developed to consider every detail in separate situations. The disposition of perceiving regularities, even though it facilitates the development of knowledge about causal connections between regularities, hinders us from perceiving any situation in its full detail.

## *4.3 Classification*

Pattern recognition is based on our ability to *classify* elements of events (Hayek 1952, s. 2.32–2.38). The ability to discern recurrent patterns arises from our ability to create categories of recurring elements in dissimilar events. The individual does not respond to separate situations as unique events (in absolute terms), but instead tries to classify their elements into certain types, based on the similarities that she can discern between the elements of the situation at hand and the categories accumulated by experience. Each perception is influenced by previous classifications. A new event is always perceived in association with the accumulated structure of elements with which it has something in common. If the elements of an event had no relation to any of the accumulated classes, the individual would remain unable to perceive the event in the first place. 'If sensory perception must be regarded as an act of classification, what we perceive can never be unique properties of individual objects but always only properties which the objects have in common with other objects' (Hayek 1952, 142).

What is assumed to happen during classification and re-classification processes is also important. Everything we perceive is related to previously accumulated classes. But also, any event contains the potential to create new and change existing classes. (Re)classification is thus a process where new events intertwine with existing categories. In order for the mind to be able to perceive order, the accumulated classes must influence perception of a new event more than the other way around. If this were not generally so, new events would continuously break down existing structure of classes and the individual would lose the ability to perceive order.

Another interesting feature in the categorising process is the feedback mechanism between the individual and her environment. Environment is generally seems as providing the feedback to which the individual then adjusts her behaviour. The individual's learning process is based on the method of trial and error (Hayek 1967, Popper 1972) where trials are hypotheses drawn upon experience and their selection is based on partly *subjective* evaluation of their respective successes or failures to achieve what is aimed for. What separates the present approach to learning from an alternative interpretation of trial and error is that not only are the trials viewed as representing the individual's subjective conjectures about causal connections, but also that the disclosing consequences are interpreted by the individual, and as the individual can perceive reality only through her subjective understanding, the degree of success or failure remains partly a subjective matter as well. This interpretation may have slightly different implications than an interpretation according to which real events work as the *objective* selection mechanism, discriminating between success and failure irrespective of the individual's assessment.

### *4.4 Multiple classification*

Classification is not necessarily a simple and straight forward process, however. An event may consist of elements that belong to more than one class at a time and they may also on different occasions be assigned to different classes depending on the accompanying elements (Hayek 1952, 50). Classification may thus be *multiple* in these two separate ways. Furthermore, classification may take place in sequences across different levels of the hierarchy of classes. One classification act may in turn become a subject to further classification, and so on (ibid., 51).

Hayek's theory of mind suggests that experience is essential for any formation of perception, that perception is essentially a process of classification of recurrent elements. The behavioural disposition of rule following is thus present in the very elementary processes by which we make the world intelligible to us.

# 5 Psychological Regularities

So far, three rationales for rule following have been discussed as genuine behavioural alternatives for rational choice theory: 1) the average (on-balance) superiority of outcomes, 2) the cost saving aspect, and 3) the cognitive limitations explanation.

The explanations that rule following brings about more advantageous consequences or that it reduces decision-making costs are problematic since it is unclear to what extent they differ from the explanation already provided by rational choice theory. Cognitive limitations seem to provide an introductory explanation that corresponds with empirical findings and does not suffer from the irrefutability tendencies of functional explanations. Whereas the other two explanations rest upon the rationality postulate, cognitive limitations suggest a departure from the logic of choice. This explanation is, contrary to the other two, testable.

A concept of regularity or rule is where our perception begins. The classification process provides an explanation of how we perceive the world, about our ability to discover similarities among elements in dissimilar events. But the fact that classification processes are multiple complicates things. If elements of an event belong to more than one class at the same time and provide different meanings when combined with different other elements, an individual's responses to slightly dissimilar events may vary. Therefore, the classifying process provides the principle on how we come to perceive regularities and it also already hints toward an explanation for why and how certain rules become socially shared. Classification presupposes a criterion (criteria) of selection by which the process becomes systematic.

Schlicht (1998, s. 7.1-7.3) recognises that the notion of a rule as an ahistorical concept does not provide any guide for future action and is therefore self-contradictory. If the set of possible rules is unlimited and there is nothing to indicate prominence, a choice among rules remains random and does not provide any behavioural indication. Schlicht criticises approaches, like that of Hayek's (1952), that try to explain behavioural principles by deriving them from associations and elementary sensory impulses (Schlicht 1998, 90 fn.). Schlicht's critique may be justified. However, Hayek offers a solution to this infinite regress problem by resorting to the evolution of the human brain, which has developed some hard-wired rules that are beyond our cognition (1952). This idea comes close to what Richard Dawkins has labelled the 'selfish gene' explanation (1978).

## *5.1 Clarity*

A prominent argument for rule following comes from our ability to recognise things that are clear and prominent in some way (cf. Schelling 1960). Clarity may manifest itself as simplicity. A simple rule may stand out

from among others and, therefore, be perceived as prominent. Another, strategic explanation would be that individuals perceive, through introspection, that introducing a complex rule would require cognitive capabilities from the part of others that are simply not realistic to expect. Complex rules have the deficiency of being open to erroneous interpretations. Axelrod's iterated Prisoner's Dilemma game demonstrates that, in social interaction, the members need not merely evaluate their actions in relation to a stable environment, but they need to evaluate the influence of their actions on other people's interpretations. Already in a two-player game a more complex rule resulted in less preferable outcomes than a simple rule, due to an increasing number of misinterpretations (Axelrod 1984, 120-1).

Individuals not only recognise clear cases, but also tend to repeat actions that have resulted in good outcomes in the past. This tendency for reproducing good outcomes has been recognised by many advocates of rule-individualism as learning by experience (e.g., Hayek 1948, 46). Another term that represents this tendency is conservatism (Schlicht 1998, 94). Conservatism describes the tendency to repeat any behavioural pattern that has previously resulted in good outcomes, whereas learning from experience is a more dynamic concept and requires the ability to cumulate knowledge as time passes.

### *5.2 Commitment*

Empirical findings show that commitment influences behaviour (see, e.g., Zimbardo and Leippe 1991). A mere decision to do something can motivate the individual to proceed without further evaluation along the way. Another, related phenomenon is the individual's inclination to finish something that she has started (for an account of classic studies of this 'Zeigarnik effect', see Koffka 1935, 334-42). Individuals seem to have a preference to 'maintain a pattern of behaviour once they have adopted it' (Schlicht 1998, 108).

Commitment may sometimes be related to obedience and authority. Stanley Milgram (1974) has made classical experiments on obedience and authority of which here is an illustrative example (found also in Schlicht 1998, 109-11 and in Zimbardo and Leippe 1991, 65-76): subjects were requested to participate in an experiment on memory and learning. The subjects are arranged into pairs and are then explained the course of the experiment by the experimenter. The aim of the experiment is to study the effect of punishment on learning. One of the subjects will act as a teacher whereas the other will be her learner. The teacher is to read word pairs to the learner and then test the learner's memory by giving the first word of each pair and asking for the word that goes with it. Incorrect answers are punished by giving an electric shock. The teacher is to push a button that releases the shock. Correct answers are not rewarded (by other any means

other than not giving an electric shock). The teacher is advised to increase the voltage after each erroneous answer, starting from 15 volts and going up to 450 volts. The buttons in the 195-240 volt range are labelled with 'very strong shock', in the 375-420 volt range with 'danger: very severe shock', and in the 435-450 volt range simply with 'XXX'.

The experiment starts and the learner receives mild electric shocks after the first few errors. Then when the voltage is increased to the level of 75 volts, the learner starts to moan. At 150 volts he starts to protest and demands discontinuing the experiment. Above 300 volts, he screams in agony, and above 330 volts he does not react any longer.

The findings of the experiment show that the majority of the 'teachers' were prepared to increase the punishment up to the maximum shock. The 'teachers' were of course the real subjects of the experiment and the 'learners' were actors. Electric shocks were not administered at all in reality, but the actors did a very good job making the 'teachers' believe the shocks were real.

The experiment used several psychological mechanisms to enforce compliance. One of the major mechanisms was commitment which grew by the preparations to the experiment. The subjects had first to respond to a newspaper advertisement and were then recruited to the experiment. They committed themselves to obeying the instructions of the experimenter which 'led them to do things that they would otherwise have refrained from doing freely' (Schlicht 1998, 110). Whenever the 'teachers' requested to stop the experiment, they would receive a standard answer from the experimenter, such as 'It is absolutely necessary that you continue', or ' You have no other choice, you must go on'. The subjects were objectively free to leave the experiment, but only a minority of them did.

Another motivating factor to continue the experiment was the gradual increase of voltage that made no particular point more prominent than others to discontinue. The teachers were also unable to predict whether or not the next answer would be erroneous, so they could hope for a correct answer in order not to be forced to administer a shock.

The Milgram experiment illustrates the power of situational factors in determining behaviour. 'Obedience in this instance was seen to override straight-forward utility-maximising behaviour' (Schlicht 1998, 110). If we consider the experiment from the subject's point of view, the outcome is perhaps not so unexpected. The subject is a layman invited to a scientific experiment. He has every reason to believe he does not possess the expertise to evaluate a scientific experiment and has, therefore, to rely on the experimenter. This is a common response to situations where the individual is already pronounced a subject to command. Another question is, however, what kinds of things people are willing to do by command. It is reasonable to expect that using a scientific experiment as a camouflage influences the outcome. Compare the discussed experiment with one where a company's CEO turns to her secretary (who is the subject of the experiment), hands

her a gun and demands her to go and shoot all the product managers of a certain business unit. Although receiving a bullet or an electric shock of 450 volts may have an equal effect on the target, it is reasonable to expect that this imaginary experiment would not persuade the subjects to the same degree as Milgram's experiment did.

### *5.3 Endowment effect*

Individuals seem to have a preference for what is already theirs. Kahneman et al. (1990) illustrate the endowment effect by an experiment where the subjects (44 students in an advanced undergraduate law and economics class at Cornell University) were offered a choice between a mug worth $6 at the nearby bookstore and a sum of money. Half of the subjects were given a mug and were assigned potential sellers, the other half were not given a mug and were assigned potential buyers. Then both the sellers and the buyers were handed a list of prices (ranging from $0.25 to $8.75 in steps of $0.50) to determine at what price they would be willing to buy or sell a mug. A real market environment was created as those price offers that met the market clearing price (which was announced after the offers were made) resulted in an exchange. The parties were offered a chance to learn from experience as four rounds of bids were administered. It turned out that the sellers valued the mugs much more than the buyers. The median reservation price for the buyers was $2.75, whereas for the sellers it was $5.25. Although 11 trades were expected to happen[2], only 1 to 4 in each round actually occurred.

There are many potential affecting factors other than the endowment effect that might explain the result. The effect of transaction costs was excluded by conducting three preceding rounds with induced-value tokens. In those rounds, 12, 11 and 10 trades were made respectively at the market clearing price. Another explanation could be that the buyers offered such low prices because they really did not need a mug. If they had, they would have bought a mug at the school's shop earlier. And the sellers made such high price offers knowing that the price level at the school's shop is common knowledge and were convinced of being able later to sell the mug at a slightly lower price than the official price at the shop. The two following experiments show that this explanation does not hold, though.

Kahneman et al. (1990) made experiments to assess the weight of reluctance to buy and reluctance to sell of a similar good as in the above experiment. A total of 77 students were assigned into three groups: buyers, sellers and choosers. The roles of the buyers and sellers were the same as in the above experiment. Choosers were asked to choose, at each of the price levels, between a mug and cash. The result again revealed substantial undertrading, only three trades took place (out of 12.5 expected ones). The

[2] Which would manifest a neutral distribution of values between a mug and money.

median valuations were $7.12 for sellers, $2.00 for buyers, and $3.12 for choosers. The outcome indicates that the sellers were reluctant to part with their entitlements[3].

Another experiment by Knetsch (1989) establishes the same effect, but this time in exchanges between two goods. Participants in three classes were offered a choice between the same two goods, a mug and a bar of Swiss chocolate. The students in one class were given a mug at the beginning of the session as compensation for completing a small questionnaire. After completing the task, they were offered to exchange the mug for a bar of Swiss chocolate. The students in another class were endowed with chocolate bars and offered the opportunity to make the opposite exchange. The student in a third class were simply offered a choice between a mug and a chocolate bar in the beginning of the session. The proportion of students selecting the mug was 89% in the first class (N=76), 56% in the second class (N=55) and only 10% in the third class (N=87). The result indicates that the possible slight income effect in the Kahneman et al. (1990) experiment does not explain loss aversion. Nor does it explain a possible valuation of the good as a type of 'trophy' as all the members in every class were endowed with a good (Kahneman et al. 1990, 1342).

Schlicht suggests an extension to the Kahneman et al. (1990) experiment: if a subject is given a mug with an understanding that she may later trade it for money and then, unexpectedly, the mug is taken away by the experimenter and the subject is instead offered a choice between the mug and a sum of money at that instant, the subject may reveal an increased willingness to pay in order to regain the mug. The subject reveals her *rule preference* which cannot be accounted for by loss aversion as the endowment effect should be weakened by taking the mug away (Schlicht 1998, 115). This experiment indicates that the moment of the *reference point change* is decisive for valuation. The subject already took the mug as her entitlement the moment the experimenter gave it to her and, therefore, the abrupt 'change of mind' of the experimenter creates a departure from the new reference point. This would then indicate, contrary to Schlicht's reasoning, that the subject's increased willingness to regain the mug reveals the same kind of loss aversion as in the other experiments discussed here.

### *5.4 Reciprocity*

A tendency to return favours or reciprocate in other ways (including retaliation) appears to be a recurring behavioural pattern among people. The degree of reciprocal behaviour varies among groups and situations, but it is a phenomenon that can be considered a regularity or a rule.

Reciprocity appears in social settings where individuals interact repeatedly. People tend to reinforce desirable and punish undesirable

[3] The positions of sellers and choosers were strictly identical and therefore the entitlement of sellers could only explain the divergence in value between them.

behaviour. Reciprocity works as a spontaneous enforcement mechanism that encourages cooperative behaviour among individuals.

Axelrod (1984) showed in his computer experiments on Prisoner's Dilemma type interaction situations that the simple strategy of tit for tat, a strategy of cooperating in the first move and then reciprocating the opponent, produced better outcomes than any other strategy. The success of tit for tat results from its combination of being cooperative on the one hand and being retaliatory on the other. The willingness to cooperate allows the actor to realise gains from cooperation in interaction situations, and being prepared to punish defection protects against recurring exploitation.

Reciprocal behaviour cannot always be explained by the narrow interpretation of self-interest. Individuals may be willing to inflict costs upon themselves knowing that the benefits resulting from the chosen course of action will not meet the costs. This is exemplified by what Trivers (1971, 49) calls 'moralistic aggression'. As punishing others inflicts costs for the retaliator, it would be expected that retaliation was resorted to in situations where there is an expectation of a future reencounter which gave a rationale for incurring the costs of the aggression. But this is not always the case. Individuals seem to be willing to use moralistic aggression also in one-shot situations where the effects of retaliation on the defector's future behaviour will not outweigh the costs of punishing (Vanberg 1994, 67).

Although moralistic aggression may, as a first approximation, seem irrational, there may be explanations to be found that give it more rational grounds. An evolutionary explanation, for instance, might suggest that adhering to a rule to always retaliate defectors, irrespective of the balance of costs and benefits, would be rational as it would provide the aggressor better protection from others' potential exploitation. A reputation of being one who is willing to inflict cost upon herself only to get the satisfaction of taking revenge may be a valuable asset (Vanberg 1994, 67-8).

A problem with evolutionary explanations is that everything can be explained assuming that regularities are always functional in some way (e.g., optimal in being suboptimal where optimality, in an objective sense, would be an unattainable state of affairs). A psychological explanation for moralistic aggression might suggest that individuals retaliate with force precisely because of the expectation of the lack of future interaction. A *preference for fairness* would trigger the willingness to incur costs because the aggressor has reason to believe that the defector would otherwise go unharmed. This explanation indicates a behavioural regularity which contradicts economic assessment as the behavioural rule. The net costs increase in situations where no future gain can be expected.

### *5.5 Loss aversion and reference level*

Literature on psychological experiments exists that explains that individuals are more sensitive to the changes in some levels or points of

reference than to the changes in the absolute characteristics of situations (Helson 1964). Loss aversion is an example of this type of phenomena. As the present situation is a prominent reference point for the individual, it is conceivable that a change from the present characteristics were perceived more intensively than a similar change in absolute terms further away. It is easier to recognise a minor change in the room temperature than an equal change from, say, 5 º C.

Loss aversion implies that people prefer the *status quo* for a situation where the probability of gain should outweigh the probability of a loss. For instance, people may prefer the *status quo* of zero gain to a 50/50 bet of winning 15 or losing 10. Risk aversion shows that the individual prefers a smaller but secure gain to a larger but uncertain one, again in a situation where objective utility would favour the risky alternative. But loss aversion suggests something different than risk aversion. At the reference point there is a kink in the utility function to demonstrate that people dislike even small-scale risk (Rabin 1998, 14). This is to say that people's attitude toward gain and loss is not symmetrical. They may be willing to forego a considerable opportunity to gain if there is any loss involved. Loss aversion may partly explain why the entrepreneur is a rare species.

Rabin (1998) suggests an interesting connection between loss aversion and reference level. The reference level explanation suggests that the individual is less sensitive to an increase in her wealth further away from her present situation. This is in line with the theory of marginal utility, the millionth pound is valued less than the first one. But, according to the reference level explanation, individuals should become risk-lovers when losses are concerned. The reference level explanation is symmetric as it does not make a difference in which direction we move from a reference point. The only decisive criterion is the distance from the reference point. A loss of one pound should therefore be more painful than an increase of loss from a million to million-and-one.

Rabin refers to Kahneman and Tversky's (1979) experiment to prove the tendency for risk preference. In the experiment, the subjects were offered a choice between (A) ) probability to loose nothing and # probability to loose $6.000, and (B) ! probability to loose nothing, but # probability to loose $2.000 and # probability to loose $4.000. 70% of the subjects preferred A. The concavity assumption of the utility function is violated as the preferred bet in the experiment is a 'mean-preserving spread of the less-preferred bet' (Rabin 1998, 15).

The reference level explanation shows that the majority of the subjects were more sensitive to the # probability to loose $2.000 as the 'third step' away from the *status quo* than to exactly the same element moved to the 'fourth step' (figure 1). It is predictable that if the values of the experiment were positive, the B alternative would dominate as the # probability to gain $2.000 somehow feels more valuable as the third probability element (the

total probability to gain something being closer to the reference point) than as combined with the fourth one.

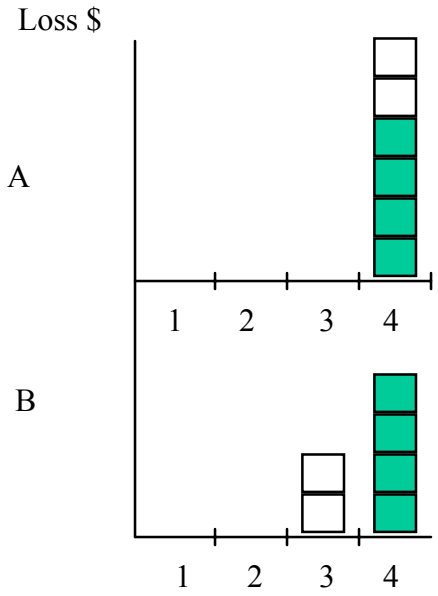

Figure 1: Reference level

In Kahneman and Tversky's experiment, both loss aversion and diminishing sensitivity seem to favour alternative A. The ) probability to preserve the *status quo* outweighs the ! probability significantly[4]. It may well be that the majority would have preferred A also in a case were the potential loss in A had been higher than $6.000. The change from the # probability to the ! probability is 100% and, therefore, the probability to loose, even a small amount, seems to increase extensively from A to B. Loss aversion would explain why the subjects 'over-valued' the probability to preserve the *status quo*[5]. The reference level explanation should show an effect in the same direction: the subjects disfavour the probability to loose $2.000 as a separate probability (moving the total probability of a loss closer to the reference point of *status quo*), thus they took it less painfully when it was attached to a less probable loss (at a point further away from the reference point).

[4] The highest significance can be calculated as the change in probability to loose anything, which is 100% from A to B (in A 25% and in B 50%). The next severe interpretation is the change in the *status quo*, which is 50% from B to A (in B 50% and in A 75%). It is important to notice that the severity of experienced change depends on the way the individual compares the alternatives. The *status quo* change from A to B is only 33% whereas it was 50% from B to A.

[5] Loss aversion is a better concept than *status quo* preference in the present use because in a similar bet with positive numbers, the *status quo* preference would not explain choice behaviour.

## *5.6 Customary rule formation*

This section discusses the question of how individuals come to adopt particular rules of custom. An important question arises about whether rules emerge from experience or whether individuals experiment with alternative configurations and select the best ones from the set. Schlicht has adopted the term 'rule inductivism' to represent the former view and 'rule structuralism' for the latter (Schlicht 1999, 3). As will be suggested below, neither view can explain the formation of rules completely.

### Rule inductivism

Customary rules become established through learning. They spread as individuals imitate recurrent behaviour of others, successful behaviour maintained and unsuccessful behaviour avoided. Rule inductivism holds that rules are formed through induction from experience. It is consistent with an evolutionary view which suggests that we use our prior knowledge to guide our future actions. The problem with the inductive approach lies in that it does not explain the formation of the initial rule (in a theoretical situation where no rule of custom yet exists). Experience is of no help if the individual cannot delimit the range of possible alternatives. A simple coordination rule of whether to pass from the right or from the left requires that the members share an understanding of what is meant by right and left. They must first have adopted a common classification system, which in turn gives rise to a second order coordination problem.

Rule inductivism, it seems, does not solve what Schlicht criticises Hayek of failing to solve. That is, it does not resolve the infinite regress problem of rules.

### Rule structuralism

This view holds that rules become selected through competition among them and, therefore, some rules must logically precede competition (Schlicht 1999, 5). This approach is closely related to views such as Vanberg's which hold that general rules, rather than separate actions, are selected according to the consequences they produce. A central question is to what extent the processes of rule formation and change can be regarded as competitive. Once a coordination rule is selected, it remains as long as it is advantageous for the people who observe it. The rule becomes a *status quo* against which other alternatives are compared. A central hypothesis of this study is that customary rule formation alleviates the competitive selection among rules. Potential rules are not constantly compared with the *status quo* due to 1) the cognitive limitations and 2) the *status quo* preference of individuals. Individuals are often ignorant about when and how to deviate (Heiner 1983). If rule following is partly explained through our cognitive

limitations to evaluate separate events, then our ability to evaluate different rules must be constrained by an even deeper ignorance. Therefore, the concept of competition does not necessarily describe well the processes of customary rule formation and change.

**Prominence and introspection**

Without the ideas of simplicity, clarity and going concern it would be difficult to imagine separate individuals arriving at similar classification systems that would facilitate coordinated action. Some rules are more prominent than others. A rule of first possession may or may not be efficient (see, e.g., Sugden 1989). Nevertheless, it stands out somehow among other potential alternatives. It is simple and neutral in the sense that when applied in an unknown case where we do not know the particularities of a situation, such as the wealth of the finder, it seems justified that the finder and not, for instance, a randomly chosen third party, receives the ownership.

Clarity and simplicity have an important communicational aspect. Without full communication, members of a group might have difficulties in interpreting and predicting actions of others. Individuals may make successful interpretations and predictions about others actions through introspection, that is, through imagining themselves in the position of others. This requires that the members share the same conventions and common experience, though. As was shown in Axelrod's (1984) computer experiments, complex rules are more vulnerable to erroneous interpretations and are dominated by simpler rules.

# 6 Procedural interest in precommitment and social learning

None of the foregoing approaches seems to provide a completely satisfactory explanation for rule following. Rational choice theory and praxeology do not view the individual as a rule-following actor to start with. Rule-individualism views choices among rules based on the same consequential assessment that is generally viewed as directing choices within rules. Categorising in Hayek's theory of mind explains the formation of perception as a categorising activity, but it does not differentiate between rule following and discretion at the observable behavioural level. Studies in experimental psychology and economics examine various observable dispositions that qualify as rules (in the sense that they are regular behavioural patterns). These experiments, however, do not address the question of how the individual comes to follow these rules in the first place.

The aim of this section is to discuss two explanations by which the individual comes to follow rules in the first place and by which the choice among rules is facilitated. It is suggested that insofar as rational contemplation is involved in the choice among rules, it may sometimes owe more to the procedural than to consequential interests. That is, a rule is chosen based on its correspondence with the set of rules that is already in place, rather than through a comparative assessment of their consequences that are yet to unfold.

Another type of process is also generally missing when rule following is examined, namely, the social learning process by which the individual internalises the procedural interest. The conclusion drawn from the discussion has two important implications, one for the examination of what kinds of games people play (in chapter 3), the other for the structural analysis of social rules (in chapter 4).

## *6.1 Precommitment*

### Precommitment based on consequential interests

In his book *Ulysses and Sirens* (1979), Jon Elster suggests that individuals are not always fully rational. In the legend Odyssey, Ulysses, the king of Ithaca, has a potential dilemma during his journey. On the one hand, he would like to hear the call of the sirens, but on the other hand, he knows that nobody, after having heard their call, has been able to resist it and has thus been doomed to their spell for all eternity. Ulysses is aware of his limits of rationality and therefore designs a procedure that binds him (both literally and conceptually) to forego the undesirable action that otherwise would

result in. He demands his crew to tie him to the mast and to block their ears so they are unable to hear his later orders.

Elster (p. 39-46) provides the principles of precommitment of this type as follows:

1. To bind oneself is to carry out a certain decision at time t1 in order to increase the probability that one will carry out another decision at time t2.
2. If the act a the earlier time has the effect of inducing a change in the set of options that will be available at the later time, then this does not count as binding oneself if the new feasible set includes the old one.
3. The effect of carrying out the decision at t1 must be to set up some causal process in the external world.
4. The resistance against carrying out the decision at t1 must be smaller than the resistance that would have opposed the carrying out of the decision at t2 had the decision at t1 not intervened.
5. The act of binding oneself must be an act of commission, not of omission.

Principle 3 disregards types of decisions that do not have behavioural effects, like decisions to decide. According to principle 5, the fact that the individual prefers not to leave a given state is not viewed as evidence that she would freely have entered that state from all of the states that are open to her (this principle has also implications to the discussion on the constitutional theory of the firm in chapter 7).

This type of precommitment can be viewed as rational within the framework suggesting that individuals are not fully rational. If individuals were fully rational precommitment would be unnecessary as the individual would be able to resist the later temptation to go against her 'true' preferences. Precommitment of this type is based on the consequential assessment of alternatives. The actor already knows what to expect to happen is she fails to precommit herself.

**Precommitment based on procedural interests**

There seems to be another type of precommitment going on in the choice behaviour as well. It was suggested earlier in this chapter that a central problem in evaluating different rules is that the individual often remains unable to assess the consequential goodness of alternative rules because rules are difficult to assess by reference to outcomes that do not yet exist. The evolutionary view on rules (which will be discussed in chapters 3 and 4) suggests that rule assessment is essentially retrospective. Only

afterwards can we assess whether a rule produced types of outcomes that we prefer. But even then, we necessarily lack the knowledge of general outcomes of other rules that were available at the time made the choice.

In the presence of ignorance about the comparative consequential efficiency of various behavioural alternatives the individual predictably resorts to the type of assessment that relates more to the consideration of consistency of behaviour, that is, to the procedural assessment. Instead of asking what rule provides, on balance, the best average consequences, the individual may ask herself what would be the proper rule to apply in this situation, and how to interpret its behavioural recommendation. This is to say that a choice among rules that can be seen as rational contemplation may be based on the individual's interest in the procedural justification.

### *6.2 Precommitment through social learning*

Due to epistemological problems concerning what we can know about rules and their outcomes, precommitment as a 'meta' rule is difficult to see as arising without social learning and interplay. Rawls' (1971) social-psychological construction provides for the possibility of a shared set of values and conventions to emerge. Rawls suggests that individuals who have a sense of themselves as individuals, and for whom pluralism with respect to final ends among all individuals is the rule, the only means to arrive at a social contract is through their *sense* of justice. If a contract is to have expected behavioural effect the individuals need to commit themselves to follow the agreed terms. But before individuals are willing to invest in a discourse leading to a potential contract they need to have expectations on reciprocity by others. By learning to precommit the individual establishes a quasi-stable reference point making her, to some extent, unwilling to defect even when defection would result in more desirable outcomes.

Chapter 8 of Rawls' *A Theory of Justice* explores how and under what conditions a sense of reciprocity arises from more primitive affections. The analysis builds upon psychological theories about stages in the child's development of moral attitudes. These theories suggest that sentiments of love and friendship, and the *sense of justice itself*, emerge from the experience of other people acting for our good. As a result of the learning process by which the child comes to recognise that others wish her well, she becomes precommitted to reciprocate in kind. There is no reason to assume that repetition could not have a habitualising effect on precommitment. There may be much truth in saying that people who deceive and lie mostly hurt themselves (that is, by the sheer acts of deceiving and lying).

This precommitment counterbalances the suggested rational disposition to unilaterally defect while others cooperate. Precommitment as a quasi-stable reference point is tolerant towards experiences of defection by others. Precommitment is seen here as a more deeply rooted regularity than what can be considered a rational strategy in the reciprocal game of tit for tat

(Axelrod 1984) which, as such, provides good reason to cooperate as well. Insofar as not all people defect all the time, habitual precommitment may explain why people are willing to endure defection, contributing to a quasi-stable social order.

Precommitment is not here seen as a universally stable pattern of behaviour, though. Even though it is expected that parents rather universally care for their children, suggesting that precommitment might be a universal regularity, the norms and institutions of a society contribute to how a precommitment at early age becomes modified as a social pattern. A further analysis of the interplay between rather universal values (such as caring for the offspring, the right to oneself, and other 'natural rights') and group-dependent institutions might suggest that non-cooperative games can be transformed into cooperative ones easier than what is expected in the rational choice framework. If individuals in general habitualise a cooperative pattern of behaviour during her childhood, it may be possible to revive that pattern in an organisation. It may be that people play the games that they assume they are supposed to play. If organisation members are seen as self-interested opportunists by the managers and social scientists, they act accordingly. It is beyond the limits of this study to pursue this line of thought further, though. Empirical evidence probably exists of the type of processes where a company that was earlier considered a hostile employer was able to revise its institutional framework and revive cooperative and trustworthy interrelations among its members. Part of the explanation in such cases may have to do with our precommitment to cooperative forms of behaviour.

Imperfect knowledge seems to contribute to the persistence of precommitment as a behavioural regularity as well. Through introspection the individual has reason to believe that other people suffer from limited reason, just like she does. The concept of defection is normally related to conflicting interest among individuals. In the Prisoner's Dilemma game (which will be discussed later in this thesis), a player defects because she expects to gain by not cooperating. She has perfect knowledge about the outcomes of cooperation and defection, for both parties. But *observed* defection in real life may result not only from conflicting interests but also from differences in knowledge about an array of activities that is considered cooperative and another that is not (this argument is developed further in later chapters). Individuals may remain tolerant toward defection if they have reason to assume that defection may be due to lack of knowledge rather than due to opportunism. Precommitment may play an important role here as well in forming expectations. If there is no convincing evidence to assume defection by opportunistic attitude, the default interpretation is that it must have occurred due to the lack of knowledge.

# 7 Conclusions

It seems that a behavioural theory that limits its inquiry to the pure logic of choice cannot provide an explanation for the formation of social phenomena, such as rules, institutions and organisations. For instance, the mixed-motives game of Prisoner's Dilemma does not explain *per se* whether or not a cooperative pattern is reached and maintained. That is to say that the pay-off structure and the basic assumption that individuals prefer better to worse do not suffice in providing an explanation of the emergence of a general behavioural pattern. As soon as other behavioural assumptions are introduced into the game, they seem to carry most of the explanatory burden. Assumptions of how much the players value e.g., the continuity of relations, trust, and reputation are pivotal to the outcome.

Psychological experiments suggest that there are numerous regularities that influence choice behaviour in relevant and systematic ways. These regularities can be seen as being based on dispositions or preferences for rule following. Loss aversion explains why individuals attach greater value to losses than to an objectively similar proportion of gains.

The predictive power of rational choice theory could be tested here. Loss aversion is generally replaced with risk aversion. But as risk is about the uncertainty indistinguishing asymmetry between win and loss, it is not a substitute for loss aversion. The central point is that one cannot arrive at loss aversion starting from the rational choice model – the direction of reasoning needs to be the opposite. What one may do, as is often done, is to claim afterwards, as empirical findings in other fields have been established, that the result fits into the boundedly rational choice model. This is a way to justify the rational choice model since empirical facts cannot be deducted from an axiom. At the individual level, rational choice theory does not exactly 'predict' much.

Reference point considerations indicate diminishing marginal utility and combined with loss aversion they indicate a kink in the indifference curve at the reference point. Rule preference indicates that individuals expect the *status quo* to remain unchanged unless there is relevant information at hand that overweighs the good continuity. Rule preference also implies that the individual's model of reality is influenced by considerations of linearity v nonlinearity and symmetry v asymmetry.

Rule-individualism explains the individual's choice behaviour from the consequential perspective. The result is a second order rational choice among rules or an emphasis on the cognitive limitations of the individual. But since we can observe that individuals engage in situational judgement as well, the cognitive limitations do not seem to explain rule following alone. Even though the individual's cognitive capacity is limited, she uses that capacity to develop expectations of the consequences that alternative choice options provide. This chapter has argued that a central issue in

differentiating between rule following and discretion are the interests that can be directed either toward consequences or toward the appropriateness of behaviour regarding the rules that are judged as beneficial by the actor.

Drawing upon the discussion in this chapter efficiency claims for or against case-by-case calculation or rule following are viewed as being problematic. If we can observe one type of behaviour the other type of response will necessarily be missing. To say that due to our limited reason rule following is an efficient response to genuine uncertainty is seen problematic as individuals do engage in action that can be described as case-by-case calculation. If case-by-case adjustment is refuted by reference to, for instance, Hayek's theory of mind which concludes that all kinds of action is based on the categorisation activity of the mind, then we can conclude that all types of action is rule following. But that would be a relabelling issue then.

Some like to think that in a highly uncertain environment, rules need to be flexible or broad in order to provide room for proper adjustment to sudden situational changes. The problem with this popular view is that humans seem to behave in exactly the opposite way. In an increasingly uncertain environment we tend to delimit the range of behavioural alternatives (Heiner 1983, Dosi et al. 1999). We apply more rigid and narrow behavioural rules when things get volatile. The reason is rather obvious: if both the environmental factors and the range of possible response alternatives became unlimited, we would lose our means of orientation.

If efficiency here refers to proper response to environmental change, then both the popular view and the one suggested here may be treated as efficient. By increasing the number of possible adjustment alternatives a more flexible or broader rule would permit a 'correct' adjustment to take place, therefore qualifying the rule as efficient. And contrastingly, by limiting the range of possible response alternatives a more restrictive rule is an efficient response to the increase in environmental volatility. What is efficiently eliminated is the risk of a response that might bring about harmful or fatal consequences. Both these views appeal to our immediate intuition. However, there is a factor which differentiates between these views, namely, empirical evidence. It can be argued that people tend to resort to increasingly simple and clear rules in an increasingly complex and volatile environment. Thus the unpopular alternative seems to be the efficient alternative. This leads to an interesting efficiency consideration: if the unpopular alternative is the general, empirically tested response, and yet the majority holds the popular view, which alternative is The efficient one?

Vanberg's version of rule-individualism appears beneficial in that it emphasises that a choice is always interconnected with the sequence of past choices. It argues that a choice is essentially a historical phenomenon, not something ahistorical and unconnected as viewed in the rational choice theory. The approach chosen here builds upon Vanberg's version on rule-individualism and suggests that a choice about which rule to follow is not

limited to the consequential assessment. Insofar as a choice among rules requires interpretation, it is the interpretation based on the procedural interest that can explain which rule is applied in a particular situation.

The rationale for the conjecture about the presence of non-consequential interests emerges from the inherent inconsistency of rule-individualism. If the cognitive limitations explain rule following, then what explains situational judgement? If cognitive limitations do not explain rule following, and a choice among rules is based on rational expectations of the comparative consequences of both rule following and situational judgement, the assumption about human capacity becomes unrealistic. Neither version seems satisfactory.

A weakness of the procedural interests explanation is that it is difficult to imagine choice behaviour that is not directed toward consequences of *some* sort. Searching for a proper rule and a proper interpretation of that rule in the particular context can be said to be consequential in the sense that the outcome that is aimed at is the correspondence between action and the rule that is considered proper. It might deserve mentioning that the notion of procedural interest does not aim at rejecting the idea of purposeful behaviour. The essential point remains that procedural interests describe preferences that give rise to a different type of choice behaviour than in consequential reasoning. A rational precommitment to long-term expectations at the cost of a short-term alternative is part of our daily decision-making. Empirical findings suggest, however, that human beings are equipped with a strong *status quo* preference that manifests itself at all levels of rules, from the personal to the social. In her book *The March of Folly*, Barbara Tuchman examines historical incidents which brought about destruction, even though anyone with any sense at all would have easily been able to see what was happening. Procedural interests are consistent with *status quo* preferences but are not limited to them. The emphasis on the interpretative aspect of the search for a proper rule and its proper application may require mental exercise that exceeds the status quo preference explanation.

# Chapter 3

# Conventions

# 1 Introduction

> Writing about custom makes me feel like a fish reasoning how water rules the life of fish. It is a staggering task. (Schlicht 1998)

Conventions are a special type of social rules. What distinguishes a convention from other regularities of behaviour among the members of groups is that almost every individual's reason for conforming to regularity includes her awareness and expectations of general conformity. Conventions can emerge either spontaneously, as an unintended consequence of the interaction among agents, or they may emerge through agreement among the participants. A shared denominator for conventions is that they provide mutual expectations of the behaviour of those who are affected by them.

Conventions will be analysed here regarding two types of rules, coordination rules and Prisoner's Dilemma rules. Lewis (1969) defined conventions as social rules that deal solely with coordination problems, such as on which side of the road to drive. Hume provides a slightly different perspective to conventions. For Hume, conventions comprise PD rules, but his analysis on the dynamics of such rules differs from those discussed in the contemporary economic literature. My aim here is to analyse Hume's position regarding conventions and PD rules. A central conclusion of the discussion that follows is that the contemporary economic literature may have adopted an unnecessarily conflict-oriented perspective to PD rules leading to an overemphasised assumption of their instability and lack of behavioural influences.

The chapter proceeds as follows: section 2 analyses coordination rules. A central issue with coordination rules is how prominence is viewed. The perspective of the present study emphasises the interpretation element in prominence. This is because if interpretation problems are seen as being already resolved in prominence, then the concept becomes rather empty: a prominent convention is prominent because it is prominent. Section 3 discusses Prisoner's Dilemma (PD) rules, and argues that the stability of PD rules is established by mutual expectations, just like in social contract. If devices need to be established to ensure stability, such devices can satisfy the mutual benefit argument. Government can thus emerge spontaneously and can be seen as part of the spontaneous stabilising mechanisms of PD rules. Section 4 analyses the affinity between PD and coordination conventions. There is good reason to assume that a transformation from PD into coordination games is an important part of social interaction. Camerer and Knez (1997) provide some interesting insight into this issue. In section 5 Hume's account of conventions will be discussed. Gauthier's (1998) analysis of Hume's approach shows close affinity between social contract and

unstable PD conventions. Section 6 discusses efficiency considerations regarding conventions. In section 7 some conclusions are discussed.

My aim in this chapter is to argue that social contract cannot resolve instability problems of PD rules. If something, social contract is an outcome, an end result, of a process by which the stability of PD is to be somehow resolved. The present study views it insufficient to say that a third party enforces PD rules. Within a normative individualistic framework, the presence of a third party must be part of the agreement. Thus, although a third party may act as the vehicle in enforcing contracts, it is not the underlying explanation for the stability of PD rules.

*Two interpretations of convention.* There are two distinct interpretations of conventions that are central to this study. The first alternative, provided by Lewis (1969), has become a generally accepted interpretation in economic literature. According to this interpretation, PD rules do not qualify as conventions. The second alternative, provided by Hume deviates from that of Lewis' in that PD rules provide the central dynamics of what Hume calls *contractual conventions*. Gauthier (1998) provides a slight modification to Lewis' version thus representing Hume's position. Lewis' definition of conventions is the following:

> 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 it is common knowledge in P that, in any instance of S among members of P,
> 1. everyone conforms to R;
> 2. everyone expects everyone else to conform to R;
> 3. everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a coordination equilibrium in S' (Lewis 1969, 58, emphasis added).

Gauthier's account of Hume's position proposes to regard

> a convention as a regularity R in the behaviour of persons P in situations S, such that part of the reason that most of these persons conform to R in S is that it is common knowledge (among P) that most persons conform to R in S and that most persons expect most (other) persons to conform R in S (1998, 19, emphasis added).

As will be analysed in this chapter the differences between Lewis' account on conventions and that of Hume's provide the central divide between what is called the orthodox and the heterodox perspectives to PD rules here.

# 2 Coordination Rules

Coordination problems correspond to individuals' knowledge problems. They can be conceptually distinguished from problems that arise from conflicting interests. We can also distinguish between an individual's preferences over constitutional environments, i.e., preferences over others' adherence to different sets of rules, and her own adherence to these rules (Vanberg 1994, 21–2). These two components do not have to cohere. A thief may prefer to live in a society where other members do not steal. Her constitutional preference is directed toward a peaceful environment where the probability of being attacked is low whereas her own adherence to the rule of private property may be called into question.

Coordination rules are special in that an individual's constitutional and compliance interests cohere (this does not imply that constitutional interests could not vary across individuals, though). The problem is not how to get people motivated to enforce a coordination rule, it is how to coordinate upon a common rule. Coordination rules are thus generally self-enforcing and do not require external sanctions to ensure adherence to them. After a coordination rule has been established and has become common knowledge, individuals have no difficulties in following it. It is the emergence and change of such a rule that provokes interesting questions.

## *2.1 Coordination problems*

A coordination problem differs from a PD problem in that it is in the interest of each participant to find a common solution, any solution. Think about a situation where A and B want to meet each other but have failed to communicate about the location (see, e.g., Lewis 1969, 5). They will meet each other only if they are able to go to the same place. If they both go to the same place, it does not matter much where it is situated. And if they fail to go to the same place, it does not matter where precisely they went as they failed to coordinate anyway. This can be depicted by a matrix (figure 3.1).

| | | B | | |
|---|---|---|---|---|
| | | C1 | C2 | C3 |
| A | R1 | 1, 1 | 0, 0 | 0, 0 |
| | R2 | 0, 0 | 1, 1 | 0, 0 |
| | R3 | 0, 0 | 0, 0 | 1, 1 |

Figure 3.1: Coordination game

The player A faces a choice among rows R1–R3 and B among columns C1–C3. Insofar as they are able to arrive at a concerted solution (any of the combinations C1–R1, C2–R2 or C3–R3) they are indifferent about which of the combinations is chosen. Any other combination is dominated by a mutual choice. A choice between which side of the road to drive, before it has become an established convention, does not supposedly arouse strong feelings for or against either alternative. But it is evidently of great benefit to everyone to concertedly arrive at either one.

Coordination problems do not have to be as 'pure' as in the above example though – purity measured by the degree of coincidence of *interests* among the participants. It may well be that A and B not only want to meet each other, but that they also have preference orderings among the places they consider relevant (figure 3.2, in Ullmann-Margalit 1977, 82).

| | | B | | |
|---|---|---|---|---|
| | | C1 | C2 | C3 |
| A | R1 | 6, 5 | 0, 0 | 0, 0 |
| | R2 | 0, 0 | 4, 4 | 0, 0 |
| | R3 | 0, 0 | 0, 0 | 5, 6 |

Figure 3.2: Coordination game with unidentical preferences

Although A would prefer R1-C1 and B would prefer C3-R3, they both would prefer either of them to the R2–C2 alternative which in turn would dominate the remaining alternatives. The central point remains: any of the coordination equilibria dominates any of the non-coordination alternatives. And more importantly, any differences among individual interests are overweighed by mutual interest in arriving at some one of the three alternatives (Ullmann-Margalit 1977, 82).

## *2.2 Prominence as a spontaneous solution*

Lewis and Schelling have argued that coordination problems are often solved spontaneously through prominence (Schelling 1960, 68; Lewis 1969, 36). Consider the following Schelling's (1960, 55–6) experiment where 41 subjects were offered the following task: 'Circle one of the numbers listed in the line below. You win if you all succeed in circling the same number.'

7 100 13 261 99 555

The first three numbers were given 37 votes, the number 7 led 100 by a slight margin, with the number 13 in the third place. The distribution of

the first two focal points was quite even. The number 7 is prominent because it is the first in the series and 100 is a power of ten.

This experiment shows that different individuals may use different conventions to arrive at a prominent solution. They need to solve a multi-level prominence problem. The first problem is to decide what type of prominence convention others might consider relevant. The convention may depend on 'analogy, precedent, accidental arrangement, symmetry, aesthetic or geometric configuration, casuistic reasoning, and who the parties are and what they know about each other' (Schelling 1960, 57). The second problem is to decide upon which alternative stands out given the chosen convention. There may exist several conventions that point toward a single alternative and also a single convention may point toward more than one alternative. Therefore, a choice among conventions may turn out to be a choice among various combinations of conventions.

In the above experiment, both the first number in the series, i.e., the number 7, and 100 are prominent depending on the chosen rule. The rule for the number 7's prominence is (here suggested to be) based on the combination of the prominence of the number one and the linear form of the series. Since the number of objects in the series is even, there is no alternative in the centre that would stand out (and therefore the rule of symmetry is not prominent). If the number 7 were replaced with the number 10, it might be expected that it gained some additional prominence because then both the prominence of the number one and the factor-of-ten rule would favour that alternative. The interpretation of rules of prominence is not always unambiguous. It is quite difficult to predict whether a randomly chosen subject will give priority to the number 10 or 100. The number 10 is prominent because human beings have 10 fingers (the primitive portable calculator). On the other hand, in the modern world, the number 100 does not only represent itself but also the 100 per cent which connects us again to the prominent convention of the number one.

Thus the rules of prominence are not necessarily unambiguous. Different rules can be applied to a single coordination problem. The degree of shared knowledge about the rules is crucial to the successful resolution of a coordination problem. There may not be many economists in the world today who would not get together at the Grand Central Station if they had to spontaneously coordinate their meeting place in New York, this of course after Schelling's (1960, 55-6) example became a classic.

# 3 Prisoner's Dilemma Rules

Prisoner's Dilemma rules differ from coordination rules in that a deviator does not immediately harm herself by the deviation, unlike if one chooses to drive on the opposite side of the road, for instance. In fact, deviation may, at first sight, seem quite a desirable mode of behaviour. Think about the rule of keeping promises. If everyone in the community keeps their promises, one can improve one's immediate position by not keeping them. Because there are incentives for individuals to violate the Prisoner's Dilemma rules, they need to be sanctioned in some way. We can distinguish between three types of sanctions: *formal* sanctions enforcement by some agency, *informal* alternatives enforced by the social environment and *internal* sanctions by, e.g., some divine entity (Vanberg 1994, 42). This section will deal mostly with the informal ones, as my intent is to discuss *spontaneous* resolutions to various problems with rules. Let us start with the basic model that shows the dynamics of the Prisoner's Dilemma situations.

## *3.1 Prisoner's Dilemma situations*

**The story of Prisoner's Dilemma**

Two prisoners are interrogated separately about a crime they committed together. There is not enough incriminating evidence to convict them without at least one of them confessing. If they both keep silent, they will be convicted of a minor offence, about which there is enough evidence against them, and each will be sentenced to one year in prison. If both of them confess, each will be sentenced to five years in prison. The prosecutor, knowing that she needs at least one of the prisoners to confess, offers a deal to each of them. The deal is that if the prisoner confesses, she is set free while the other accomplice will be sentenced to ten years in prison. This setting can be demonstrated using a simple game-theoretical matrix (figure 3.3).

| | | B: Not confess | B: Confess |
|---|---|---|---|
| A | Not confess | 1, 1 | 0, 10 |
| | Confess | 10, 0 | 5, 5 |

Figure 3.3: Prisoner's Dilemma

The rise of the dilemma is due to the strategic features of the situation. The prisoners are separated from each other and cannot therefore communicate. If they were able to communicate, they could strengthen their mutual trust relation and strike a deal to keep silent. Another important factor is that each of them is fully aware of the rules of the game. The pay-offs are obvious (the inverse of the number of years in prison) and each of them can anticipate that the same offer to walk free is made not only to her but also to the other party. This gives rise to strategic considerations.

Each prisoner must choose between keeping silent and confessing without knowing in advance the other party's choice. If A was informed in advance that B did not confess, A would be tempted to confess and thus walk free. On the other hand, if A was informed that B already confessed, she would certainly also confess. This dynamics indicates that prior knowledge about the other party's choice does not affect one's choice behaviour in this simple game. This is to say that the outcome would be unaltered if the prisoners were to choose in turns so that the choice of the first chooser would be explicit to the second one.

The choice to confess dominates the choice not to in this game. This dominance is due to the structure of the pay-offs and the strategic relation between the prisoners: 'if A confesses, his pay-off is higher than it would have been had he decided not to confess, regardless of B's choice of action' (Ullmann-Margalit 1977, 19) (Being freed dominates one year of imprisonment and five years of imprisonment dominates ten years).

The dominance of confession leads to the Nash equilibrium, being Pareto-inferior. It is an equilibrium in the sense that each prisoner would be worse off if she chose to unilaterally deviate from it. Suboptimality is due to the fact that both would gain if they could jointly move to the not-confess mode of behaviour. But since a unilateral move leaves the chooser vulnerable to exploitation, a rational chooser will forego that alternative.

### A generalised PD structure

The above story of a dilemma of two prisoners is instructive as it reveals *non-arbitrary* pay-offs of the prisoners. Although the actual years in prison may vary from case to case, the structure of the pay-offs remains unaltered. The dynamics of the game is thus not the result of *a theorist* arbitrarily setting the pay-offs for the players.

The game between two players can be extended to comprise any number of individuals, firms or other groups of people. The reason why the game is usually demonstrated with using two individuals has to do with considerations of clarity and simplicity. The dynamics of the game is easier to recognise through a simplified, clear model. The game can thus be generalised as follows (See, e.g., Ullmann-Margalit 1977, 23): a PD situation is any situation involving at least two individuals each of whom faces a choice between cooperation or defection under the following conditions:

- If all of them defect, the outcome is (and is known to them to be) mutually harmful.
- If all of them cooperate, the outcome is (and is known to them to be) mutually more desirable than if everyone defects.
- Each participant would derive the highest pay-off by defecting while all others cooperate.
- One's defection while the others cooperate is, at least partly, at their expense. That is, the others would gain if the single defector would choose to cooperate.

In real life contexts, the dynamics of the game can be effective without as severe a polarisation as in the above general model. All can sometimes be replaced with most and instead of one defector we can assume there to be more than one who defect without influencing the choice behaviour of those who cooperate. These features have to do with *knowledge* and *tolerance* in a group. In a group where most pay taxes an additional tax evader gains while the others may either be ignorant about the number of tax evaders or simply tolerate their behaviour. This is to say that although the tax payers may not know the actual degree of harm the evaders inflict, or may not care about it, they are in an *objective* sense made worse off.

This connects us to the theme whether somebody can be made worse or better off without her knowing of it, or, to put it in more general terms, should we approach values and utilities as objective or subjective matters. A Pareto-optimal state is generally referred to as one in which it is not possible to improve the position of at least one individual in a group without harming anyone else (upper left-hand corner in the PD game). This interpretation leaves open the question of who is to decide upon the changes in values. Pareto efficiency can be analysed both from the subjectivist and from the objectivist perspectives. This is an understandable consequence since Pareto efficiency *per se,* no more than the PD game, provides answers to the question about sources of valuation.

**Solutions to PD problems by aligning consequential interests**

It is clear that PD dynamics lead to undesirable outcomes in a group, that is, undesirable to all compared to another outcome that could only be reached through cooperation. A central question then arises about methodology to facilitate cooperation. One prominent solution to guarantee compliance is to influence the pay-off structure that each member faces. Ullmann-Margalit (1977, 30–3) provides a good example of the method. Imagine two mortarmen in two isolated outposts facing an enemy attack. Each of them faces a choice between remaining at his post and fighting back, or fleeing from the post. The dynamics of the game is the following: (1) if both stay and fight, they are able to shell the enemy and repel the attack. (2) If both flee their posts, the enemy will break through and take both of them prisoner. (3) If one stays at his post while the other flees, the

one who stays manages to hold the enemy up just long enough for the other to escape safely, but is killed as the enemy will eventually break through. Assume that both mortarmen are aware of the dynamics of their interrelated situations. The pay-offs are illustrated below (figure 3.4).

| | | B: R | B: D |
|---|---|---|---|
| A | Remain | 1, 1 | 2, -2 |
| | Desert | -2, 2 | -1, -1 |

Figure 3.4: Pay-offs of the mortarmen

Fleeing while the other stays and fights back gives the highest pay-off as (it is here assumed that) fighting is more costly than deserting. But if both choose to flee, the outcome for both is less desirable than if they both had stayed and fought. So here we have an apparent PD situation.

Our purpose is then to change the pay-off structure in order to facilitate a situation where both mortarmen loose the incentive to flee. One potential solution, presented by Ullman-Margalit (1977, 32), would be to lay mines around the posts of the mortarmen. This would change the perceived pay-offs to the following pattern (figure 3.5).

| | | B: R | B: D |
|---|---|---|---|
| A | Remain | 1, 1 | - 2, -2 |
| | Desert | -2, -2 | -2, -2 |

Figure 3.5: Pay-offs of the mortarmen surrounded by a minefield

Thus, any attempt to flee from the post would result in a guaranteed death. The mortarmen would have an increased incentive to stay and fight until the last breath.

The above solution being a slightly brutal attempt to stabilise the cooperative mode of behaviour one can come up with other solutions that do not perhaps give as guaranteed outcomes but have the same tendency to alleviate defection. Ullmann-Margalit (1977, 33–7) discusses *discipline* (see also Sen 1974, 59ff.) and *honour* as alternative stabilising devices. A central difference between these two solutions and minelaying is that the two former are *norms* whereas the latter can be viewed as a unique action to solve the PD situation at hand. Minelaying can of course become a norm if it is applied in every similar situation. The conceptual distinction between a

designed single solution and a norm is instructive, so let us assume that minelaying represents the former.

The two mortarmen may belong to a unit where military discipline is severe. This discipline may comprise a rule that defection under enemy attack will lead to execution. If we compare this rule to minelaying, some differences can be found. In order to be executed a defector has to be caught first. The uncertainty of execution may provide an incentive for a mortarman to flee if a battle situation looks hopeless or if he realises that the other mortarman is fleeing. The pay-offs are as follows (figure 3.6, also in Ullmann-Margalit 1977, 33).

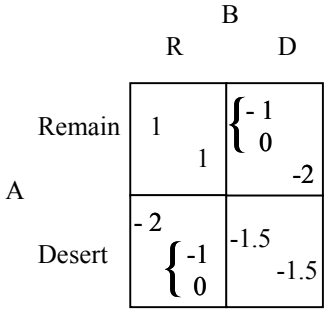


Figure 3.6: Pay-offs of the mortarmen constrained by discipline

Remaining in the post results in the same pay-off (1) as in the figure 2.2. Also, staying and fighting when the other deserts gives an unaltered pay-off (-2). To desert when the other deserts too gives a pay-off of -1.5 which is more desirable than staying behind and getting killed but less desirable than the initial pay-off resulting from captivity. This pay-off can be constructed by combining the (un)desirability resulting from captivity and the probability of being later executed. To desert when the other stays behind gives either –1 or 0 depending upon the probability of being caught by own troops and hence executed.

The point Ullmann-Margalit (1977, 33–6) makes by using this pay-off structure is that in this version, there are two equilibria: (R, R) since –1 and 0 are worse than 1 and (D, D) since –2 is worse than –1.5. This means that neither mortarman finds remaining in or deserting from the post rational *irrespective* of the other's course of action. The worst that can happen when remaining in the post is –2 while the worst pay-off is slightly better (–1.5) if one chooses to desert. On the other hand, remaining results in a higher pay-off (1) than deserting (-1 or 0) when the other remains too. This indicates that each mortarman will desert if he suspects the other to desert but as long as he expects the other to remain his maximising choice will be to remain also.

The dynamics of interdependency in this version has some important features that are worth considering more generally. The model illustrates

open-ended environments where individuals are constrained by various threats of punishment, the probabilities of which, however, remain uncertain and subject to individual judgement.

A central idea of Ullmann-Margalit is to show that in PD-type problems 'certain devices (generally speaking, PD norms) will be generated' (ibid., 35) to secure the cooperative mode of behaviour. For her, a PD norm is one that 'when supported by sufficiently severe sanctions, is capable of solving, or indeed dissolving, potential generalized PD-structured problem.' (ibid., 38). 'The efficacy of this norm [of prohibiting desertion in battle], conjoined with the severe punishment for its breach, solves the problem inherent in this type of situations' (ibid., 40).

The type of PD norm that Ullmann-Margalit discusses here as the solution to PD situations is one that is *purposefully designed* to solve the incentive problems inherent in the situation. Military discipline is a designed tool to facilitate obedience among soldiers. The norm to execute those who desert the battle has also rather straight forward behavioural consequences. Not all the norms that can be seen as solving PD problems need to be deliberately designed, however.

In what follows, my interest is to investigate the relation between conventions as spontaneous *solutions* to recurrent PD situations and recurrent PD situations as *problems* that need to be solved. This interest is connected to a popular idea that PD rules are efficient solutions for PD problems (as, e.g., in Ullmann-Margalit 1977, 40). This idea is decomposed into two separate issues. The first question is, to what extent we can consider PD rules as solutions to PD problems? The second issue is about the efficiency claim. The question then is about the method with which we should be able to evaluate efficiency.

### *3.2 PD norms as solutions to PD situations*

This section focuses on spontaneous rules that solve PD problems as unintended consequences. The central idea of these rules is that while individuals act self-interestedly, they produce socially beneficial outcomes that enhance cooperation among actors. I will use the term *PD norm* to refer to these rules.

An important characteristic of PD norms is that they require action that is socially beneficial but that may be in conflict with the immediate interests of the actor herself (Vanberg 1994, 42). An individual's commitment to keeping promises benefits others while in some situations it may turn out to be against the immediate interests of the individual to do so. Another important feature of PD norms is that it is in everyone's indirect interest to maintain a PD norm. An individual prefers a state of affairs where all participants, herself included, adheres to a PD rule in another state where the rule does not exist. This implies that the benefits from, e.g., the

trustworthiness of others exceed the costs of having to keep promises herself.

**Are all PD norms moral rules?**

What is said so far about PD norms can be said about moral rules. In fact, a special feature of moral rules is that they contain the PD problem. To behave morally means to abide by a PD rule whose destruction would be harmful to everyone. But the original question was whether or not all PD norms are simultaneously moral rules.

According to Gauthier (1967), a system of moral rules is one that is constituted of principles that influence individuals' choice behaviour to enhance cooperative action in PD situations. As a consequence individuals are willing to forego their maximising choice (to defect) in order to maintain the perceived benefits from cooperation. Thus, a moral person will not defect, not even in a situation where she can anticipate the other party being unable to retaliate later by defecting too.

Ullmann-Margalit (1977) challenges this definition of moral rules. She refers to an example where two gangsters behave 'morally' (toward each other) by not confessing although it would be in each party's interest to defect. She questions whether the outcome would suffice for 'them to be considered moral men?' (ibid., 42) Ullmann-Margalit recognises that Gauthier's interpretation would give an answer in the affirmative whereas her own position is in the opposite. A further condition is suggested by her: an outcome should not involve 'disadvantage to anyone extraneous to the PD-structured situation under consideration' (ibid.).

Two things are worth noting about this claim. First, it is true that Gauthier's positive interpretation does not provide a normative judgement about what particular rules ought to be considered moral rules. The two gangsters may act morally toward each other by deceiving others. If the two gangsters abide by the moral norms of, e.g., the mafia and consider the norms of the surrounding society partly irrelevant, then we have a problem of conflicting norms between a group and its sub-group. Norms of either unit cannot be justified without reference to the values of their respective members. There seem to be no positive means for us to differentiate between PD norms and moral norms.

The second point in Ullmann-Margalit's interpretation that needs to be discussed is about the priniciple of not harming anyone extraneous to the PD situation. It may turn out to be rather difficult to collect information about the unintended negative consequences of, e.g., an investment. We have every reason to believe that many investments in the Western countries harm people living in the second or third world countries — without any deliberate intention. These investments are, if consistent with Ullmann-Margalit's interpretation, immoral. Then we are back in the first issue about

whose values are to be used in evaluating this[6]. To illustrate the problem of evaluating harmful unintended consequences picture yourself standing on a pavement. The sheer occupation of that space can inflict harm on others who are walking by. As an extreme interpretation, being alive would be immoral.

Insofar as PD norms are considered *exogenous* to a PD problem, there is not much difficulty in showing how the problem is resolved. If the pay-off structure is altered by the adherence to moral rules, a cooperative mode of behaviour can be stabilised. What makes PD-type social norms interesting as solutions is their *endogenous* dynamics. A spontaneous rule that solves a PD problem comprises the same PD problem in itself. Although the rule of keeping promises solves PD problems by promoting trust, it is vulnerable to exploitation precisely because of its function.

## *3.3 The rise of PD rules*

### The state-of-nature approach

As was noted above, institutions can be analysed as exogenous constraints to behaviour. Then it is interesting to examine what influences the existing or theoretically constructed institutions may have on behavioural patterns. Another approach that will be advocated here is to analyse the rise of social institutions from the behavioural patterns themselves. This approach is advocated by theorists who are primarily interested in explaining the rise of social institutions as unintended consequences, that is, as spontaneous processes of interaction among individuals (see, e.g., Hayek 1973, Ullmann-Margalit 1977, Schotter 1981, Sugden 1986).

The central idea of this approach is to start the analysis in a state of nature where no social institutions yet exist and individuals are only equipped with their preferences and skills (Schotter 1981, 20). The aim is then to provide an explanation or a reconstruction of how social institutions *could have* risen from the initial state. This does not mean that the institutions we can observe today, such as government, necessarily have emerged according to an evolutionary story provided by a theorist. It only shows that under the conditions that can be reasonably constructed, certain types of institutions can emerge without deliberate design. The usefulness of the explanation depends upon the assumptions about the initial conditions. My aim is to analyse the nature of self-interest in connection with the rise of the institutions discussed below.

---

[6] Ullmann-Margalit's position is not remedied by limiting the notion of harm to refer only to *intended* harm. The two gangsters do not necessarily intend any harm to other people. As an unintended consequence, though, some people may get hurt by their ways of doing business.

Robert Nozick (1974) has used the state-of-nature approach to demonstrate how the state can arise in a non-coercive way through the interaction among self-interested individuals. Locke defines the state of nature of individuals' as 'a state of perfect freedom to order their actions, and dispose of their possessions and persons as they think fit, within the bounds of the law of Nature, without asking leave or depending upon the will of any other man' (Locke 1986 [1690], s. 4). This state is, however, not anarchy in the sense of absence of *rules* although it can be viewed as anarchy in the sense of absence of *government*. This is presented in Locke's normative statement that 'no one ought to harm another in his life, health, liberty, or possessions' (Ibid., s. 6).

Nozick (1974) demonstrates how protective associations can arise spontaneously in the state of nature. An individual member of a group may get other members to join in in her defence or retaliation against a defector because of various reasons: the members may be public spirited, or they may be her friends, or she may have helped some of them before, or they may wish her to help them in the future, or they may act in exchange for something (ibid., 12). This indicates that in the state of nature, individuals adhere to moral rules, they are also willing to help other members either because they value friendship or because of the prospective gains from reciprocal behaviour and reputation. This has an important indication contrary to Schotter's above interpretation of the point of departure. In the state of nature, individuals are not only equipped with their preferences and skills. In order for a protective agency to spontaneously arise, individuals need to have an understanding of the potential gains of reciprocal behaviour when there are expectations for recurrent interaction.

Thomas Hobbes' (1996 [1654]) interpretation of the state of nature differs from that of Locke. In Hobbes' model, individuals are in a deeper state of anarchy, that is, in a 'war of everyone against everyone' (p, 84). A central difference between these models is that in Hobbes' state of nature, individuals do not respect others' natural rights to their life nor to their property. The state of war is considered stable because it is in nobody's interest to act as the sole peace-keeper. The one who alone pursues peace and performs his promises, when nobody else followes suit, will 'but make himself a prey to others, and procure his own certain ruin' (p. 103). On the other hand, according to Hobbes, in the state of nature, it is in the interest of individuals to keep agreements with those who keep agreements with others (ibid., ch. 15). This introduces a logical problem. If it is in nobody's interest to start cooperating when nobody else has yet done so, how can individuals exist with cooperative reputations?

Hobbes' model shows an important aspect to the resolution of PD problems. It is precisely his assumption of the self-destructiveness of being the first mover that prevents any possibility for cooperation to emerge. There is only one way to solve the problem of the first mover within the model, and that is by introducing reciprocity and reputation. If they are part

of the behavioural assumptions, then the assumptions in Hobbes' model do not apply anymore.

*The rise of moral rules.* As I argued above, Nozick's illustration of the state of nature is not, contrary to Schotter's interpretation, a state where no rules yet exist. What Nozick explicitly demonstrates in his analysis is how the group members can resolve PD problems by relationships where the expectation of future cooperation is valuable.

The rise of moral rules is consistent with the assumption that individuals are self-interested. There is no reason to believe that individuals will not consider gains from reciprocal behaviour and from their reputations as relevant elements in their preference structure. A narrow version of self-interest can be demonstrated by the PD game where the players are trapped in the Hobbesian *status quo* of defection. In order for the players to be able to improve their situations, they have to discover trust and commitment. In the Hobbesian model, recurrent encounters among the players do not facilitate an improvement as each player is fixated on the first mover's dilemma. In the Lockean model, recurrent interaction is central to the development of reciprocity and reputation. A subtle but important difference between these models appears to be that the Lockean model equips the individual with the capacity of introspection. Cooperation is possible only insofar as the individual is able to put herself in the position of others. That is to say that without introspection there is no reciprocity and without reciprocity there is no cooperation.

This approach emphasises the characteristics of moral rules as arising and changing through endogenous processes among interdependent individuals. Solutions to PD problems, whether spontaneous or designed, presuppose some shared expectation, some initial *convention*. Otherwise, an evolutionary story or a reconstruction of the initial conditions would not show any development toward the kinds of orders we can observe around us. A war of everyone against everyone else would have no end.

# 4 Prisoner's Dilemma rules as conventions

This section focuses on a conjecture that analyses of conventions based on game-theoretic models have overemphasised the distinction between cooperative and non-cooperative elements in human interaction. The fact that we can conceptually distinguish between these elements does not prove that real interaction follows the strict dichotomy of these concepts. The dynamics of the coordination game suggest that the individual's choice of which type of action to take in a coordination game is a social decision; it requires the individual to think about what others will do and what shared rules they adhere to.

Camerer and Knez (1997, 166-7) argue that under three common conditions, games that are normally classified as PDs are essentially coordination games because players prefer to reciprocate cooperation. The first condition refers to the individual's internal representation of the game as a problem of coordinating the level of cooperativeness. In many one-shot PD games, the players often forego the purportedly maximising alternative of defection and cooperate instead (cf. Rabin 1993, Ledyard 1995, Sally 1995). The second condition refers to situations where there is synergy or complementarity in the cooperative behaviour. This refers to cases where the individual is a member of a team whose produce becomes valuable through the mutual contribution of the team members; where the participation in the group *per se* is valuable for the member; and where expectations of network externalities and critical mass are present. Also, easy identification of defection contributes to the cooperative mode of behaviour. The third condition refers to PD games that are played repeatedly with a quasi-infinite time horizon among players who are not excessively impatient. The Folk-theorem in game theory implies that cooperation is the wealth-maximising strategy insofar as deception triggers tit-for-tat punishment in later rounds. Thus the repeated PD game sometimes transforms into a coordination game.

It is reasonable to assume that economic organisations represent a type of social group where these dynamics influence the member's willingness to cooperate. Organisation members are not independent to design and impose their separate interests upon others without influencing the perceived goodness of the cooperative game.

### Convention and organisational expectations

The application of coordination games to organisational coordination problems emphasises the mechanisms that transform expectations of proper action to be taken rather than mechanisms that transform preferences (Camerer and Knez 1997, 172). That is to say that it is the interpretations of organisational rules, based on mutual procedural interests of the participants,

rather than conflicting consequential interests that guide organisational behaviour.

Organisational expectations presuppose common expectations of how other members will behave. March (1997) and Zhou (1997) suggest that organisational expectations support organisational norms that guide proper actions of the members in particular choice situations. It seems that we are back in the infinite regress problem of rules. Common expectations support organisational conventions, and common expectations are arrived at by conventions that point toward the particular expectations. The expectations of proper action in an economic organisation are derived from the conventions of the surrounding society whose members the organisation members assumably are. The structural consideration of rules in the constitutional analysis emphasises the view adopted here that agreement, whether explicit or implicit, becomes a central criterion of goodness.

The social learning process and precommitment as a procedural interest (as discussed in the previous chapter) become important factors in attaining common expectations. Both notions refer not only to coordination problems but also to cooperation problems. The individual learns about theories that other people have about states of affairs and causal connections (producing what can be described as objective knowledge [Popper 1972]), but she also learns to abide by the moral rules that she expects are relevant in particular situations.

Insofar as mutual expectations are central to the definition of convention, there may be not enough reason to exclude PD norms from among conventions. The rationale for including PD norms is not based on some arbitrary relabelling move. It may well be that a strict dichotomy between conventions as self-enforcing, non-conflictual coordination rules and PD norms as inherently unstable and conflictual rules requiring external enforcement misleads us to assume that the latter carries less behavioural influences on human interaction.

# 5 Hume's account on convention

The reason for conforming to a convention relates to the interest of individual members to do so, and includes both a preference for general conformity *per se*, rather than a consideration of the outcome of general conformity, and a preference for personal nonconformity unless there is general conformity (Gauthier 1998, 19). So, instead of considering consequential efficiency, the individual members direct their preferences toward procedural considerations. Any solution is acceptable insofar as it is based on shared expectations of general conformity.

The essence of convention is that personal conformity to a convention is each person's most preferred response to conformity by others, and each person's least preferred response to nonconformity. A convention is *dominant* insofar as it is not seriously dispreferred to any alternative regularity for behaviour in a similar situation, and it is *stable* insofar as conformity is not seriously dispreferred to nonconformity, given conformity by others (Gauthier 1998, 20).

If a convention is stable, the members do not need a *covenant*[7] to enforce conformity since it is in the direct interests of everyone to conform. And, if a convention is also dominant, there is no need for explicit agreement as it is obvious to all parties what the expectations of others are. A stable and dominant convention, such as on which side of the road to drive, is thus not contractarian in the sense that no contract needs to be established for its maintenance. This section focuses on the *contractual conventions*, that is, those conventions that are not dominant and/or not stable. I argue that implications drawn from the contractual conventions, especially with the PD conventions, are misleading and based on an inconsistent logic of reasoning.

Contractual conventions are characterised by two devices that bring about coordination of interests among the participants over alternative conventions and security of conformity to a convention that is established. A *bargaining* process is central to the establishment of a contractual convention because the participants need to resolute their opposed preferences in order to select a mutually beneficial alternative. I would argue that assumptions about such a bargaining process influence our view about how stable a chosen convention will be.

The basic approach in economic literature, based on the assumption of self-interest with guile, views bargaining processes as essentially competitive where each participant aims to give as little as possible to gain as much as possible. Hume's view to bargaining is different. Bargaining is

[7] Covenant is here interpreted as in Gauthier (1998, 21): 'an agreement, entered into by each person on the basis of his own interests, which assures, with or without enforcement, mutual adherence to a convention'.

essentially effected through an appeal to each individual's interests. As the participants can, through introspection, reflect upon what type of conventions are mutually considered beneficial, the process is not described as being competitive. If a participant has had false assumptions about general interests, she can learn and revise her view during the bargaining process. The aim of the participants is to find some prominent principles that are mutually shared, not to give less and gain more.

The second device to facilitate stability of contractual conventions is the covenant that was mentioned above. By covenant we mean an *agreement* that assures mutual adherence to a convention. Such an agreement may or may not include the use of a third-party enforcement. What is essential for a covenant is that adherence to a convention against one's immediate interests is assured through an appeal to interest (Gauthier 1998, 21). To understand this paradoxical argument let us look at how property rights and government are justified in Hume's terms.

**Property and government as conventions**

Gauthier (1998, 30ff) explains how for Hume, property rights are agreed upon due to two dynamics of contractual conventions. First, each participant expects to benefit from a system of property compared to having no system at all. Second, none of the participants expects to benefit much more if an alternative system was introduced (assuming the mutual benefit criterion). Thus a convention that stands out due to its salience becomes accepted.

When it comes to adherence to property rights, Hume is no idealist:

> All men are sensible of the necessity of justice to maintain peace and order, and all men are sensible of the necessity of peace and order for the maintenance of society. Yet, … such is the frailty or perverseness of our nature! it is impossible to keep men faithfully and unerringly in the paths of justice. Some extraordinary circumstances may happen, in which a man finds his interests to be more promoted by fraud or rapine, than hurt by the breach which his injustice makes in the social union. But much more frequently, he is seduced from his great and important, but distant interests, by the allurement of present, though often very frivolous temptations. The great weakness is incurable in human nature. (1987, Pt I, No. V, Of the Origin of Government)

Thus the conventions of property would be stable if individuals were guided by their overall interests rather than by their short-term temptations. The central problem with unstable contractual conventions (PD conventions) is that each participant prefers universal conformity to a convention of property, over general nonconformity, but at the same time, each prefers not

to conform in specific situations although others conform. Hume considers *obligation* and *enforcement* as supports for our real interests to maintain the stability of a property convention.

Obligation influences interests to conform as each participant reflects, not on her own failure to conform, but on the consequences of general failure (Gauthier 1998, 35). This obligation counter-balances our inclination of nonconformance. But since it is not assumable that all participants are capable of performing this balancing act, third-party enforcement may be needed.

A basic argument in economic literature is that since PD conventions are unstable, government is called to sanction defection. A conclusion drawn from this, that PD conventions thus fail in providing behavioural influences, is, I think, incorrect. Hume's train of logic of seems to be the opposite: since we have reason to assume that some PD conventions need explicit enforcement, we set up a third party to enforce such conventions thus resulting in both the stability of the convention and its behavioural influence to all participants. As Gauthier puts it, 'those who wish to secure the obedience of their fellow can best do so by "the impartial administration of justice". Hence those who exercise power tie men to obedience, and in order to exercise power, they in turn are tied to justice' (1988, 37).

The mutual interest in establishing property thus extends to the mutual interest in establishing government. Thus, instead of viewing PD conventions as failing to become established, Hume's position emphasises the ability of individual participants to precommit themselves to secure the overall benefit resulting from general conformance. Establishing government is thus analogous to Ulysses' demand of binding him to the mast (Elster 1979).

# 6 Are conventions efficient solutions to problems?

This section will analyse some implications that the distinction between consequential and procedural interests may have on efficiency considerations regarding conventions. I shall consider two main questions. The first one is about whether we can reliably know what particular problems particular conventions solve. The second one is that if the first question can be answered in the affirmative, can we consider solutions to be efficient in some sense? Answering these two questions involve both consequential and procedural considerations.

It is maintained here that due to epistemological limitations it is often impossible to establish the consequential efficiency of conventions whereas the procedural assessment of efficiency is a viable option.

Schotter (1981) takes a straightforward position to explain why certain types of institutions emerge while others do not:

> The problem facing social scientists is to infer the evolutionary problem that must have existed for the institution as we see it to have developed. Every evolutionary economic problem requires a social institution to solve it (Schotter 1981, 2).

Schotter refers to the famous example of money as a means of exchange. Money emerges as a solution for the problem of efficient multilateral trade (Schotter 1981, 3). He quite reasonably recognises that historical solutions are *indeterminate* in the sense that other solutions might have evolved had events taken another path. What can be expected though is the *type* of solution. As for money, *ex post* reasoning is a simple and straightforward process. As we already know that money exists what we need to do is to discover potential problems that it resolves. Another way to see this would be to say that conventions do not always solve problems (which are nonexistent at the time of the emergence), but rather, they facilitate development that would otherwise have been different.

Explanations for the emergence of conventions can be assessed, among other things, through their reliance on *functionalism* and *adaptationism*. A functional explanation is one that refers to something, e.g., to a social institution by its effects (Vromen 1995, 90). Money emerges because it solves the problem of efficient multilateral trade. According to this view a convention that does not solve any perceivable problem would be difficult to imagine. Adaptationism goes even further than functionalism by claiming that prevailing institutions are optimally adapted to the environment. Thus, existing conventions do not only solve problems, but they do it in the most efficient way.

If we apply functional adaptationist reasoning to predict future institutions, things do not appear so simple anymore. This is due to the epistemological problem that we do not know *ex ante* what type of institutions will emerge and prevail. The reason for that lies in the nature of spontaneously emerging institutions. These institutions are not consequences of anybody's intentional design and can be perceived only after they have emerged through the interaction among people.

Another approach to conventions as solutions, that does not presume knowledge about unintended consequences, is based on the procedural interests of the participants. According to this approach conventions resolve problems of indeterminacy in interaction among individuals in recurrent situations that have multiple equilibria (Young 1966, 105). This Humean perspective corresponds with Lewis' (1969) view that it is coordinated interaction *per se* that is the source of desirability. Individuals prefer *any* mutual solution to a coordination problem, to the lack of solution. Conventions could then be taken to satisfy the assumption of functionality in the sense that it is always beneficial to have a common coordination procedure. On the other hand, this perspective does not support the adaptationist claim that prevailing conventions should be optimal.

**Conventions of distributive bargaining**

The question, why a fair share between two who have contributed equally is so often agreed to be 50-50, is fascinating to social theorists. For instance, Skyrms (1996) provides an evolutionary explanation for the stability of 50-50 as emerging behind the 'Darwinian veil of ignorance' (p. 10). In this model, the idea is to ask pairs of individuals, who have potentially dissimilar ideas about a fair share, to share a cake with each other. If the combined percentage is over 100, they both will lose. If the combined percentage is below 100, sharing is possible. The dynamics of the model show that no matter what the initial distribution of individuals (regarding their degrees of selfishness/altruism) the natural selection tends toward the 50-50 rule.

The main issue that bothers me in this model is the assumption that if two selfish (say, 82-63) people bargain, they will both get nothing. I assume that this assumption is crucial for the model to work properly. Now I assume that this model was designed to explain the real life convention of 50-50 as a fair share. In real life, there is no reason to assume that pairs of 82-63 could not bargain and strike a deal. Neither is there reason to assume that pairs representing 21-14 shares would be willing to give 65 per cent of their combined property to a third party.

By changing few basic assumptions, the model becomes insolvable. The 50-50 rule is prominent, but for other reasons than suggested by Skyrms. I would argue that our sense of fairness has developed through our ability to perceive others as fundamentally similar human beings as ourselves

(see further in Rawls 1971). The 50-50 share is, I would argue, not based on some arbitrary equilibrium pull of 50-50, but, instead, on the prominence of the *finders-keepers* rule. If there happens to be two persons in a situation of sharing something, and there are no other prominent rules to interfere, 50-50 is a prominent outcome.

# 7 Conclusions

The basic treatment of efficiency in economic literature is viewed here as being directed to consequences. The suboptimality of conventions derives from the expectations that some other, even more desirable convention might exist but is unattainable because of the lock-in features of conventions. The desirability of a convention is assessed by the consequences that such an alternative is expected to bring about. If conventions were assessed by the procedural interests of the participants, that is, by the degree that people adhere to a convention, an existing convention that gets near perfect conformity would be assessed as efficient. This perspective is analysed further in the next chapter where the efficiency criterion of constitutional economics is examined.

Conventions of property manifest a combination of coordination and (non)cooperation features. Such conventions can be called mixed-motives conventions. The participants prefer a system of property to its absence, but on the other hand, separate participants may have differing interests as to what particular system is adhered to. The Humean perspective to conventions emphasises the procedural interests of the participants. The participants can expect that as soon as everyone starts importing their personal, consequential interests into the process of bargaining, the attainment of agreement becomes improbable.

If we model social interaction by assuming individuals' interests being directed only toward the consequences of their choices, PD situations are prevented mainly by influencing the payoff structure. In economic organisations, direct control and performance-based rewarding are central to facilitate conformity among the participants. But if the participants also have procedural interests in finding the appropriate interpretation of rules, the disregard of such interests *per se* may provoke attitudes that bring about an undesirable pattern.

The discussion in this chapter implies that to establish stability of PD rules in a community, attention directed to enforcement issues is attention directed to symptoms. Genuine stability can only be brought about by focusing on the multilateral reciprocity relations among the participants. The orthodox version of self-interest in economics may provide an unnecessarily pessimistic picture about people's ability of to play well together.

# Chapter 4

# Social Contract

# 1 Introduction

This chapter examines the principles by which purposefully designed rules are established from the normative individualist perspective. Contractarian reasoning adheres to the subjectivist ideas that individuals differ in their knowledge and interests. The contractarian principles are not only consistent with methodological individualism, which suggests that social phenomena should be examined by deriving them from the actions of individuals, but they are also founded in the notion of normative individualism, which precludes any value judgement apart from the individuals concerned.

The criteria of efficiency or goodness, that correspond with the limitations of normative individualism, will be discussed. The discussion provided in this chapter will imply that strict adherence to normative individualist principles does not allow the introduction of criteria of goodness, independent of the rules enforced by the relevant community.

Contractarian reasoning emphasises a structural analysis of rules in searching for the justification for rules through the procedures that give rise to them. An alternative, evolutionary approach is discussed in relation to contractarian analysis. From the evolutionary perspective, a unanimous agreement to enforce a rule represents the *trial* element in a process that can be specified as the trial and error (elimination) process (Popper 1979 [1972], 261; Hayek 1988, 20). The evolutionary approach addresses the inability of the contractarian model to fully capture an omnipresent application of rules, namely the unintended consequences that necessarily follow.

The chapter proceeds as follows: section 2 provides the basic principles of contractarian philosophy. In section 3 the normative foundation of contractarian reasoning is analysed. It will be argued that efficiency considerations, based on normative individualism, are limited by the normative impact of rules that are practiced in a community. Opportunity costs are considered as a potential candidate in breaking with the normative content of rules. The present study is, however, unable to find a satisfactory way to view opportunity costs as positive entities, disconnected from the rules by which they are created and changed. Section 3 will also discuss procedural and consequential issues regarding efficiency of rules. The distinction between the procedural and the consequential approach to efficiency is viewed beneficial because it clarifies a central aspect of efficiency: disregarding which perspective one is to choose, the efficiency consideration always remains a partly subjective task and thus subject to speculation in a social context.

The contractarian approach maintains that constitutional rules guide the development of rules of just conduct. The discussion at the end of this chapter will argue that another, evolutionary position is equally valid in its view that constitutional rules are themselves dependent on rules of just

conduct. A priority of the spontaneous can be justified for the following structural reason. The first step away from complete anarchy requires exchange. Insofar as exchange is of asymmetrical nature where all parties do not gain immediately but, instead, some parties need to keep promises and others need to trust that promises are kept, shared expectations are required. Conceptually, shared expectations cannot emerge through agreement, which itself requires the presence of the expectations in question. Thus, any agreement presupposes a convention of shared expectations that cannot be explained by social contract.

# 2 Contractarianism: an individualist approach to collective action

As an explanatory approach, contractarianism suggests that collective, organised arrangements are viewed as networks of multilateral, interindividual relations that are different from the market relations among participants. The essential difference is that intraorganisational relations cannot be factored down into separate bilateral exchange relations (Vanberg 1985, 21). This point of view is contrary to the nexus of contracts approach to economic organisation (Alchian and Demsetz 1972), which suggests that intraorganisational relations are indistinguishable from market relations, and can thus be mechanistically factored down into separate contracts between each member and the organisation itself. What makes, from the contractarian point of view, the decomposition of intraorganisational relations impossible is the multilateral, reciprocal exchange of commitments to the constitutional order of the organisation.

The constitution of an organisation specifies the terms of participation: (1) which resources participants are to contribute to the organisation, (2) how and by whom the decisions on the use of pooled resources are to be made, and (3) how the resulting benefits from the joint endeavour are to be shared among participants (Vanberg 1985, 22). These constitutional rules are distinguishable from rules of just conduct which define the boundaries between private realms of individuals. Constitutional rules are viewed to be structurally prior to the rules of just conduct, even though it is recognised that not all rules of just conduct require guidance from the deliberately designed constitution (ibid., 24).

## *2.1 Methodological and normative individualism*

Contractarian reasoning is based on two libertarian meta-theoretical principles, *methodological individualism* and *normative individualism* (Vanberg 1985). Methodological individualism is an explanatory principle guided by the general idea that social phenomena should be explained as the aggregate outcome of the interaction among individuals, each pursuing her private interests, under certain constraints. Normative individualism is a normative meta-theory providing a proper means to evaluate social states. Normative individualism demonstrates that, because of the necessarily *subjective* nature of knowledge and values, the relevant standard against which the goodness of social states is to be judged are the individuals involved.

Since it is recognised in normative individualism that individuals vary in what they know and what they want, both intertemporally and across individuals, an external observer cannot reliably gather information on the desires of other people in any other way than observing *actual choices* being

made (Buchanan 1977, 102). Information revealed by individuals about what they *would* want *if* circumstances were different does not fulfill the requirement.

The focus on the moment of choice brings a central implication to normative individualism: there are no evaluative criteria against which the goodness of social states can be judged as such, independently of the individual choices that resulted those outcomes (Vanberg 1985, 3). As social states are results of *processes* constrained by certain *rules*, it is those processes and rules that become central in social study. The goodness of social states is only indirectly assessed, dependent on whether (or to the degree by which) the processes and rules that brought these states reflect what the individuals involved want.

The justification for a liberal social order lies in the normative premise that individuals are the ultimate sovereigns in matters of social organisation (Buchanan 1991, 227). Individuals are entitled to choose the organisational and institutional settings under which they themselves will live. It should be emphasised that Buchanan is explicit in noticing that the sovereignty of individuals does not provide a normative legitimacy for organisational structures that allow the most extensive range of separate individual choice (ibid). That is to say that individuals are altogether entitled to choose rules and institutions that restrict their future range of choices. This perspective is therefore not applicable for advocating ideologies of unlimited freedom of choice or laissez-faire or other types which view freedom as a supraindividual value.

## *2.2 Exchange paradigm*

The constitutional perspective highlights voluntary exchange as the core motivator for the individual to limit her behaviour within constraints (Buchanan 1991, 5). The cost of limiting one's own behaviour is accepted insofar as it does not exceed the benefits expected to result from reciprocal behaviour of others. This perspective emphasises the calculative rationality of the individual who actively chooses her own constraints, if only 'to a degree and within some limits' as Buchanan adds (ibid.). The exchange paradigm involves inquiry into the cooperative arrangements of interaction among individuals (Buchanan 1991, 9). By definition, a voluntary exchange happens only when all participants gain from the trade.

A Humean account of social contract emphasises the procedural interests of the participants whose mutual goal is to attain an agreement on *any* contract that can be expected to be just, as assessed by the participants themselves. There may be a slight difference between the notions of exchange of commitment and mutual interest, the latter being possibly more tolerant toward human errors and shortcomings, thus providing more stability for PD conventions.

The notion of *implicit social contract* will be discussed in relation to the concept of exchange. Entering into and remaining in an organisation is here taken to reveal an implicit acceptance of the social contract of which the entrant has not necessarily been part. The idea of an implicit social contract is important as it remedies the possible inefficiency arising from the fact that social contracts also constrain individuals who have not explicitly agreed upon them.

### *2.3 Unanimity as the contractual ideal*

The subjectivist position of the contractarian perspective recognises that values and theories about social phenomena vary across individuals. This limits efficiency considerations because it is believed that no supraindividual scalar of goodness exists. There is no reason to believe that the ordering of preferences did not vary across individuals. The fact that individuals know different things and do not necessarily interpret events in the same way links us to the knowledge problem of society (Hayek 1945).

From the subjectivist position, an assessment of efficiency relies on revealed preferences of the individual. A voluntary exchange between two parties is taken to indicate that both parties gained from the exchange. It should be noted that the efficiency of an exchange does not carry further than to the event itself. Either party may regret the exchange if she afterwards has reason to change her view. When the idea of voluntary exchange is transferred to the realm of collective choice, the strict criterion of revealed preferences through observed exchange needs to encompass all the parties. As the subjectivist position holds that the values of individuals are incommensurable, an exclusion of any one party from the exchange breaks down the possibility to verify that the observed exchange was in fact efficient.

According to contractarian reasoning, not all choices among rules need to satisfy the strict criterion of unanimity, though. It is entirely justifiable for any group to unanimously agree upon relaxing the criterion for types of rules that are specific to the extent that a complete agreement would be too costly to achieve (Buchanan and Tullock 1962). This makes the perspective more operational but at the same time subject to infinite regression. This relates to the problem of determining the degree of unanimity that is required of a choice defining the category of the degree of unanimity a particular rule should be placed into. The participants can expect that negotiating a particular rule into a category of more strict unanimity (e.g., 1/2, 2/3, 3/5, etc.) affects its probability of becoming established, and therefore conflicts of interests are being provoked. If an agreement on the category of a particular rule is less than unanimous, it becomes itself a target for rent-seeking. This is to say that even though a unanimous agreement justifies the *principle* of enforcing rules based on less than unanimous agreements, the choices of what category to apply in

particular cases cannot escape the infinite regression problem. If an agreement about which alternative category is to be used in particular cases is not unanimous, the contractarian criterion of goodness does not apply.

### *2.4 The veil of uncertainty*

Since individuals vary in their theories and interests, it is likely that when a group of people get together in order to pursue something collectively, conflicts of interests arise and mutual agreement may thus be difficult to achieve. The members need to compromise before a mutually agreed solution can be reached. The solution perhaps does not match perfectly with anybody's personal interests but provides a more desirable outcome than what being left without it would result in. The question about how to facilitate a compromise thus becomes central. A compromise requires the parties, to some extent, to alienate their immediate self-interests and, through introspection, assess what would be expected by the other parties.

Rawls (1972) suggested an idealised normative construction of *the veil of ignorance* as the proper starting point for individuals in their pursuit for the basic principles of justice. According to the model, individuals are believed to be able to alienate themselves completely from their personal positions in their community. On the other hand, it is held that individuals possess perfect knowledge of the general facts about human society. The epistemological position advocated here is, however, limited to assumptions about the individual's reason and cognitive capacities that do not readily permit Rawls' formulation about the initial position. A concept that works in the same direction as the veil of ignorance, but in a more positive manner, is the notion of *the veil of uncertainty* (Brennan and Buchanan 1985).

Brennan and Buchanan introduced the veil of uncertainty to provide more realistic epistemic assumptions about the individual. Individuals are not completely alienated from their positions in society, or from their epistemic capabilities. On the other hand, the individuals' knowledge about the general working properties of rules remains imperfect. The point that Brennan and Buchanan (1985, 29) make is that the constitutional choice process *itself* contains aspects that facilitate agreement. Rules are by definition more general than the outcomes that result from action guided by those rules. A constitutional choice among alternative rules contains the elements of generality as a chosen rule needs to be applicable in numerous contingencies. Another basic characteristic of a rule is its extended time horizon. A rule needs to be applied over time, otherwise it can hardly be considered a rule. Due to these considerations, the individual faces genuine uncertainty about how her position will be affected by the operation of a particular rule. Insofar as mutual agreement is the goal, the individual tends to agree on rules that can be considered fair in the sense that they are broadly acceptable within the relevant community (ibid., 30).

### Theories and interests

As stated earlier, the constitutional perspective shares the assumption of the exchange paradigm that individuals participate in organisations because they expect to gain from the membership. Potential disagreement on constitutional rules may be due to two conceptually distinct components, a theory component and an interest component (Vanberg 1994, 167). The participants' perceptions about the working properties of alternative rules may vary. Although all participants might be motivated to discover fair and impartial rules, the fact that they do not interpret the working properties of rules in a similar fashion may prevent an agreement. This can be viewed as the theory component of a constitutional choice.

The social contract perspective, represented by, e.g., Rawls and Buchanan, sees the potential disagreement arising primarily from the inability to reach a compromise among the parties due to conflicting *interests*. The interest-oriented perspective directs our attention to procedures that facilitate agreement though alienating personal interests of the parties. If an agreement is reached, the criterion of goodness is the voluntariness itself. The distinction between procedural and consequential interest discussed in this study implies that to the extent that individuals are affected by their procedural interests, conflicts of interest can be alleviated.

The theory-oriented approach, represented by, e.g., Habermas (1990), emphasises the discourse process that can be compared to a scientific dialogue (cf. Popper 1995 [1945], Polanyi 1951). The emerging social contract is then seen as a discovery process during which different interpretations about the working properties of alternative rules may change or become refined to the extent that the parties share a common understanding of them. This perspective links to the Hayekian 'dispersed knowledge' problem of society (Hayek 1937). A discourse process can facilitate hitherto undiscovered alternatives, which would be preferred by all parties, to disclose. Hayek's position can justifiably be considered closely related to the interpretation of social contract as an open-ended discovery process (eg., Vanberg [1986a] specifies social contract as conjectural).

Disagreements in collective endeavour arise from both conflicting interests and dissimilar theories about states of affairs. Hayek's (1952) theory of mind suggests that conflicts of interest arise even among well-intentioned actors because of their inability to consciously direct law and resources to predefined goals in the way they believe they can. Therefore, even if actors were specified as being altruistic or rule-followers, conflicting interests would arise because of their inherent inability to know their own minds. This is to say that the notion of conflict of interest not only refers to dissimilar preferences among separate individuals but also to the fact that insofar as Hayek's theory of mind holds, individuals cannot articulate their preferences in their entirety.

# 3 An assessment of the normative premise of contractarianism

## *3.1 Efficiency through revealed choices*

An implication of normative individualism considered here holds that the only reliable way for an observer to assess efficiency of an exchange is the act of exchange itself. This view relates to the universal explanatory theories of *praxeology* (Mises 1966 [1949]) and *rational choice*. These theories provide the tautology that the individual always prefers better for worse, the logical truism which holds in every choice situation. Insofar as the logic of choice is considered to hold, the act of exchange *per se* reveals enough information for an observer to be able to conclude that the exchange benefited both parties. What the observed exchange reveals is that both parties assessed the exchange to be the best option among the perceived alternatives that were open for them at the moment of choice. The assessment begins and ends at the instance of choice (Buchanan 1969). Both praxeology and rational choice theory are silent about the assessment of consequences of exchange – and so is normative individualism.

According to contractarian principles, social states cannot be evaluated independently of the individual choices which give rise to them (Vanberg 1986). Normative emphasis is, therefore, shifted from the outcomes to the processes that give rise to these outcomes. If a particular choice among rules brings about certain general outcomes, then the goodness of those general outcomes needs to be assessed indirectly through the rules that guide the process by which the choice was made. 'Social states are only indirectly to be judged as 'good'; such judgement depends on whether the process by which they are brought about can reasonably be assumed to reflect what the individuals involved want' (Vanberg 1985, 3). This amounts to what is labelled here as procedural assessment of efficiency.

The interplay between rules and outcomes has interesting implications for normative individualism with respect to the assessment of goodness of rules. As was stated above, general outcomes can only be assessed by referring to rules that give rise to them. This is logical since rules are then viewed as causes and outcomes as effects. But when the same logic of assessment is applied to the rules themselves, things get more complicated. The contractarian point of view suggests that the goodness of rules should be assessed against the processes by which they are established. And these processes against the rules that give rise to them, and so on. This logical process leads to infinite regress (Vanberg 1986).

**Procedural-consequential goodness criterion**

An alternative way to assess the goodness of rules, which I shall develop in the following, is to accept the contractarian procedural-structural criterion of goodness, and then combine it with a consequential assessment of the degree to which the *observed* general outcomes correspond with the *expected* ones, judged by the relevant individuals. The perspective I have in mind can be labelled as *procedural-consequential* assessment of efficiency.

From the procedural-consequential perspective, the assessment of goodness is then not only dependent on the voluntary exchange aspect of agreement, but it also takes into account the *ex post* assessment of the – intended and unintended – consequences of rules. The rationale for this extension is that the 'very nature of rules implies that their "goodness" can only be judged by their performance over a longer sequence of applications' (Vanberg 1994, 29). This extended approach implies then that an agreement upon enforcing a certain rule can be specified as a *trial* and a collective response to observed outcomes can be specified either as a *corroboration* of the trial, or as its *falsification*, which then leads to a new trial. The overall process during which trials and error corrections occur can thus be referred to as a trial and error (elimination) process (Hayek 1967, Popper 1972).

Figure 4.1 illustrates the position adopted here. The procedural justification is represented by the vertical dimension where each rule is assessed against the process that gives rise to it, which in turn is assessed against the extent to which it corresponds with rules of a higher order (of generality). At t1 point in time, the goodness of the outcome (dotted line) of rule r1 is assessed against rule r1's consistency with the rule r2, which in turn against rule r3, etc. Insofar as the establishing process of rule r1 corresponds with rule r2, the consequences that rule r1 produces are justified.

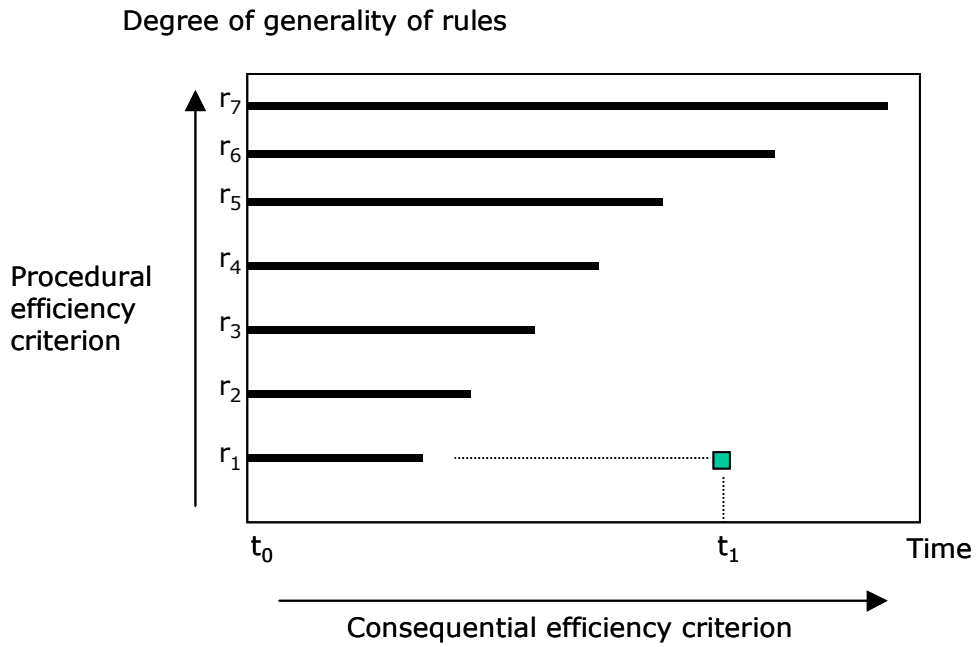


Figure 4.1: Procedural and consequential criteria of goodness

The consequential justification focuses on the outcome along the horizontal timeline. The goodness of the rule r1 is assessed against the degree of desirability of the outcome at t1, independent of the justifiability of the rule r1 itself. The consequential justification can be decomposed into *ex ante* and *ex post* considerations. The former corresponds with rational expectations, where expected outcomes direct choice behaviour. The latter assessment is possible only after the outcomes have unfolded. The question about a proper time horizon is disregarded here.

The notion 'outcome' works as a bridge between constitutional and evolutionary perspectives. When we look at the negotiation phase of a social contract, it is quite clear that an essential element that affects the agreement is the assessment of the *working properties* of alternative rules under consideration. These working properties refer to the *general expected consequences* of rules. It would become impossible to describe an agreement as a voluntary *choice* among the parties if the choosers had no idea of what the general expected consequences of alternative rules were. From the contractarian perspective, general outcomes that affect choice are those that we *expect* to appear after the rule has been enforced.

The evolutionary approach emphasises general outcomes from a different perspective. All social actions result in unintended consequences (as a minimum, knowledge is altered in an unpredictable way). The general outcomes we expected to occur may or may not occur, and irrespective of whether they occur or not, unintended consequences necessarily occur. By

definition, at the moment of choice we remain ignorant about the unintended consequences. From the evolutionary perspective then, general outcomes comprise both intended and unintended consequences.

If we go beyond the procedural-structural perspective of contractarian reasoning, it becomes possible to extend the assessment of rules to encompass unintended consequences as well. This links back to what was earlier mentioned about the cognitive element in rule assessment, namely that the goodness of rules is not necessarily only based on procedural considerations but comprises also the general observed outcomes that they result in. This clearly establishes a connection to the ability of the market process to select, not only by reference to voluntary agreement, but also by reference to the outcomes that are at the moment of agreement necessarily nonexistent.

### Implications to constitutional and evolutionary positions

The perspective I am advocating views the evolution of rules as follows: There is clearly an important component of purposeful, collective deliberation involved. However, this does not make the evolutionary process *per se* a process of purposeful selection. I think that the notion of *purposeful selection* (cf. Commons 1924) may be slightly overstated. It is correctly based on the assumption that purposeful deliberation is involved when people get together to design rules. But the fact that people get together to design *trials* which bear influence to the overall order, the nature of these influences remaining largely unknown, does not yet mean that the people involved somehow purposefully select the overall evolutionary process *per se*. The fact that after implementing a trial (a new rule) people often need to get back together to revise what they designed simply because even the preliminary experience of outcomes is disappointing, gives a rather haphazard view of the purported purposefulness of cultural evolution. To refer to cultural evolution as purposeful selection amounts to the same as seeing the market participants as purposefully selecting the overall market process. The fact that the market process, as well as cultural evolution, is *constituted* by a myriad of separate choices among rules and within rules, based on purposeful deliberation, does not make the overall process purposeful *per se*. This distinction is missing from the interpretation of cultural evolution as purposeful selection.

The perspective advocated here thus deviates from both the efficiency-driven evolutionary approach and the rationalistic-constructivist constitutional approach. The differences are subtle but important. The reason for calling purposeful selection in cultural evolution an overstatement is because it leads to an unfounded view to the overall evolutionary process as something that can be designed purposefully. The intermediate position that I am advocating here is not in any way hostile toward purposeful,

collective processes. But it should be enough to say that these collective processes produce trials into a complex overall process of evolution.

Consider an example of a community where an agreement is established on a rule prohibiting drinking in public places (excluding places that are licensed to serve alcohol). Because the rule has been enforced for such a long time, the members of the community have no experience-based knowledge about the consequences of a rule that permitted drinking in public places. Later on, the members have, however, observed that while visiting other communities where drinking is allowed in public places, no obvious problems seem to have appeared. What happens then is that the members unanimously agree upon imitating the rule that permits public drinking, based on their observations of other groups. The rule change seems to be well justified as it corresponds with a more general rule of personal freedom. After the permitting rule has been established, the members are able to observe the intended and also the unintended consequences that emerge afterwards. After a reasonably long period of time, the members come to the conclusion that, based on the newly gained experience within the community, the rule change was not for the better. So they agree again unanimously to change the rule back to prohibit drinking in public places.

From the normative individualist perspective both agreements, the permission and the prohibition, enjoy equal normative value, since they are assessed against the processes by which these rules were established. And since both rules were established by unanimous agreement, neither of them enjoys more procedural justification than the other. The procedural-consequential approach suggested here would appreciate the value of agreement *per se*, but in addition, it would also encompass the consequences that resulted from these rule changes. The conclusion would be that the prohibition of public drinking was consistent with the preferences of the members. The information needed to establish this was not available at the moment of the first agreement to permit public drinking, though.

An agreement can be established only by using knowledge that is there at the moment of agreement. Similarly, from the consequential point of view, the goodness of a rule change can only be assessed after we have gathered knowledge on the general outcomes by experience. The essential difference is the differential temporal dimension between these approaches. While the procedural-structural approach can only assess efficiency at the moment of choice, the procedural-consequential perspective is more encompassing temporally, including outcomes that begin to bear their influence on assessment only afterwards.

It is important to acknowledge, however, that the procedural-consequential approach also remains an imperfect means to evaluate efficiency. The critical factor comes from the interrelation between time and ignorance (O'Driscoll & Rizzo 1985). In the above example, it is implicitly assumed that the preferences toward personal rights of the members are

more or less stable during the whole process of the rule changes. But, of course, this need not be the case. The point that is implied in the procedural-structural view is that we cannot reliably compare a combination of a choice and its outcomes with another occurring later on. This is because time has necessarily elapsed and knowledge is not unaltered (Lachmann 1976, 127-8). The two agreements on changing the rule, first to permitting and then prohibiting public drinking, cannot, with complete reliability, be compared against each other because the knowledge inherent in the two separate choice situations is necessarily dissimilar.

Hayek's evolutionary approach tries to resolve the problem of indefiniteness of assessment of consecutive choices among rules by assuming that the evolutionary development *per se* brings about appropriate rules, not necessarily because the individuals involved understand them as being appropriate, but because the individuals involved as a group outperform other groups with inferior rules (Hayek 1973, 18f). Vanberg (1986b) maintains that there is no reason to assume that a spontaneous process, upon which we could blindly rely, could be accounted for giving rise to appropriate rules. This issue will be discussed in the next chapter.

When considering mutual benefit, it is easy to take the normative individualistic position and conclude that insofar as exchange was observed, all benefited. A closer look at the dynamics of the institutions that frame exchange may prove that a straightforward efficiency claim becomes speculative. The rules under which an exchange takes place define what can be judged as a mutual gain. The boundary between acceptable and unacceptable exchange, and between acceptable and unacceptable consequences for third parties needs to be defined *prior* to the assessment of mutual gain. Next I turn to examine to what extent the contractarian approach can contribute to specifying ways to define the boundary between voluntariness and coercion.

## *3.2 Voluntary agreement*

### Voluntariness vs. coercion

All actions that take place in a social setting have unintended consequences (Knight 1935, 53). Third parties are always somehow affected by exchange transactions[8]. In a community, the boundary between acceptable and unacceptable externalities needs to be defined. It is conceivable that since the details of transactions are unforeseeable and their general attributes may change with time, the definition of the demarcation between externalities that are tolerated and those that are not, needs to be based on a set of rules that is applicable in dissimilar situations.

[8] Conceptually, third parties are affected by the sole exclusion from the exchange transaction.

When we engage in exchange with others, unintended consequences are unavoidable. Especially interesting are consequences that affect third parties. The normative content of efficiency of the market process has to deal with how the rights of third parties are permitted to be affected. Although a voluntary exchange benefits the parties involved, it may have adverse effects on others. We have to deal with the boundary between acceptable and unacceptable consequences that third parties are potentially subject to.

There is a clearcut criterion that distinguishes between acceptable and unacceptable consequences. Insofar as the *rights* of other people are not violated, an exchange can be judged to have no adverse affects on third parties (Vanberg 1986, 119). The problem of defining the boundary between voluntariness and coercion is, however, not yet solved. If rights define the boundary between voluntariness and coercion, then the rules by which those rights emerge become a central issue. Two alternative ways to approach these rules seem to be available. Either they are taken to be absolute norms whose validity is independent of what particular rights a community adheres to, or they are viewed as dependent on the recognition and enforcement of the relevant community (ibid.).

Empirical findings show that rights vary across communities and groups. It is of course possible that the rules that give rise to particular rights are absolute but so abstract that differences in the interpretation of their meaning give rise to heterogeneity of rights across groups. But in that case those rules do not bear a normative content on how rights are to be defined. A second option is still available to us, namely the group-dependent rules. If the rules depend on social recognition and they vary across groups, there seems to be no prominent boundary in the continuum between voluntariness and coercion beyond which rights could not develop. A totalitarian system would then be, at least conceptually, morally feasible and it would still fulfill the requirement of the connection between the rules and the rights that emerge as an outcome. What we may want to do is to examine ways to specify a criterion for discriminating between voluntary and coerced choices that would be independent of the prevailing sets of rules in any particular community (Vanberg 1986, 120).

There are three potential ways to specify group-independent rules: (1) resorting to the idea of absolute rights, discussed above, (2) applying the same procedural individualistic criterion of goodness to rules that was used to the market process. From the procedural-structural point of view, this would give rise to infinite regression, however, as those interactive processes were judged as good that are based on good rules, which themselves are to be judged against a criterion of goodness applied to the process by which they emerged as an outcome, and so on. (3) The third alternative is based on opportunity cost assessment. The parties to voluntary exchange can easily, i.e., with low opportunity costs, withdraw or refrain from engaging in transaction (Hirschman 1970). A coerced choice would then be a situation

where there are high opportunity costs of avoiding or exiting from the exchange (Vanberg 1986, 122). This criterion can be exemplified in a social setting by reference to the comparison of opportunity costs of resigning from the membership of a business firm with those perceived when emigrating from one's home country. A sophisticated normative individualist position recognises that the opportunity cost reasoning alone is not sufficient to provide justification in assessing goodness. It needs to be combined with the procedural criterion that emphasises the voluntariness of exchange or agreement (ibid., 134).

Implications of the second and the third options are of interest here. The normative individualist approach holds that rather than being defined by the rules of the community, a normatively significant notion of voluntary choice has to be defined as a standard against which the rules themselves are to be assessed (Vanberg 1986, 120). Thus, the second alternative, which is vulnerable to infinite regression, does not seem to qualify. The third alternative may seem to fare well in the search for a universal, independent criterion that would not be dependent of the social rules of a community. But does it serve as an end-station to the chain of normative assessments?

### Opportunity cost as a positive criterion of goodness

Opportunity costs can be interpreted in basically two alternative ways. The first alternative is to view opportunity costs as objectively existing entities, independent of the evaluating subject. This interpretation receives justification from considering, for instance, market prices as objective entities that remain more or less unaltered irrespective of whether or not a particular agent chooses to engage in exchange. But in order for this interpretation to be acceptable, we must assume perfect competition where the market prices are in harmony with the subjective assessments of opportunity costs. Insofar as normative individualism is founded on subjectivist principles, the objectivist interpretation of opportunity costs seems not to be an available alternative, however.

The second alternative is to interpret opportunity costs as subjectively perceived valuations over those alternatives that are recognised by the chooser at the moment of choice. A strictly subjectivist interpretation of cost is provided by Buchanan (1969, 42-3):

> Cost is that which the decision-taker sacrifices or gives up when he makes a choice. It consists in his own evaluation of the enjoyment or utility that he anticipates having to forego as a result of selection among alternative courses of action. The following specific implications emerge from this choice-bound conception of cost:
>
> 1. Most importantly, cost must be borne exclusively by the decision-maker; it is not possible for cost to be shifted to or imposed on others.

2. Cost is subjective; it exists in the mind of the decision-maker and nowhere else.
3. Cost is based on anticipations; it is necessarily a forward-looking or *ex ante* concept.
4. Cost can never be realized because of the fact of choice itself: that which is given up cannot be enjoyed.
5. Cost cannot be measured by someone other than the decision-maker because there is no way that subjective experience can be directly observed.
6. Finally, cost can be dated at the moment of decision or choice.

Buchanan's interpretation on cost provides an epistemic underpinning for normative individualism. Opportunity cost is something that is perceived by the individual at the moment of choice, and it vanishes immediately when the choice is made. We can only observe the world we are living in, not imaginary worlds where the alternatives we did not choose carry their consequences.

If opportunity costs are interpreted in terms of subjectivist principles, the assessment of goodness is limited to the moment of choice in the same way that was demonstrated in the voluntary exchange model of normative individualism. Thus, even though we can observe that some individuals may reveal that they face higher opportunity costs when exiting a country than a business firm, an observer is not allowed to assume that this is the case with everyone else as well.

Irrespective of how subjective or objective opportunity costs are taken, another problem remains about their independency of rules enforced in the relevant community. In order for opportunity cost to function as a positive element in the combinatory criterion of goodness (together with the procedural criterion), it should not depend on the rules of the community. Opportunity cost would then be the element capable of breaking the problem of infinite regress. In order to function as such an element, opportunity costs should be able to be taken as given at some level of rule making. Could this, even in principle, be justifiable (even when assuming the possibility of objective opportunity costs)?

A problem with opportunity cost in the context of social contract is that a social contract influences the general opportunity cost structure among the members. The main rationale for agreeing to limit one's future behaviour is the expected reciprocal constraint on others. An agreement of this type influences the opportunity costs perceived by the members. Consider a divorce rule that requires a trial separation period of six months before a divorce becomes finalised (Elster 2000, 13). Although such a rule is not a good example of what is normally taken to be constitutional, it illustrates the point I am trying to make here. This rule may gain status of close to unanimity in a group because most rational actors can perceive the benefit of trial period in tempering passions. A rational actor is assumed

here to prefer being Ulysses bound to the mast to being Ulysses unbound. On the other hand, such a provision increases opportunity costs when entering a marriage (by making exit more costly, at least in the temporal sense). Here again the interpretation of cost becomes speculative. The provision is costly for a married couple who after the trial period fail to resolute, and may have wished to have been able to, e.g., remarry during the trial period. Obviously, the provision is less costly, or actually beneficial, for a couple who manages to resolute during the trial period.

Consider a clause in a contract between a business company and its manager, which prohibits the manager from entering any firm in the same industry within two years after the termination of the contract. The relevant criterion of goodness is still assumed to be the combination of the procedure (agreement) and the low opportunity cost. A problem with the combination in this case is that the agreement depends on the expected opportunity costs and *vice versa*. A primary goal for the company may be to prevent adverse consequences that an opportunistic manager might impose on the company by taking her firm-specific knowledge to a competitor. A rational manager should expect such preventive measures and is willing to accept the clause insofar as she is compensated for the increased opportunity costs. The agreement is thus feasible only when accompanied with high opportunity costs which are then compensated somehow.

Now a question arises whether the opportunity costs are reduced by counter-balancing them in, e.g., monetary terms. If the answer is in the affirmative, there should be no objections against terms of contract that come close to what can be taken as slavery. The point I am trying to make here is that any combination of procedural goodness and opportunity cost enjoys equal moral justification. If, on the other hand, the answer is in the negative, then increased opportunity costs cannot be justified by agreement. But then the opportunity cost element is not only combined with the procedural consideration, but in fact becomes the *primary* criterion. This implies that there is no independent way to assess the combination of procedural and opportunity cost criteria of goodness.

Another problem in opportunity cost assessment concerns whether opportunity costs are inflicted *incidentally* or *essentially* (as inferred from Elster's [2000, 4] discussion on incidental and essential constraints). An essential opportunity cost arises when the agreement is *intended* to increase opportunity costs, as is the case in the above example. Opportunity costs can be also inflicted incidentally through unintended consequences arising from choices made by others elsewhere. The easiness to emigrate depends not only on the exit rules of the community from which one is to emigrate, but also on the rules constraining immigration by other communities. Moreover, the perceived easiness to exit and enter depends on the expected opportunities that a particular community can offer in a myriad of ways (ease in finding a job, personal and social security, etc.). These considerations imply that opportunity costs depend on the various rules that

are adhered to in communities, and on the subjective contingent assessment by the chooser.

A community that is based on some interpretation of freedom needs to have some rules that define the private realms of freedom of the people. These rules themselves affect not only the subjective evaluation of opportunity costs, but also the general patterns of opportunity costs in a community. Therefore, even if the subjectivist foundation is disregarded, the general patterns of opportunity costs are dependent on enforced rules. The only way to justify the idea of opportunity costs being independent of all the rules enforced by a community would be to assume opportunity costs emerging in an institutional vacuum.

The normative individualist approach tries to solve this dilemma by referring to the idea that efficiency of voluntary exchange or agreement on a rule is *conjectural* rather than definite (Vanberg 1986, 135). This allows the possibility that parties to exchange might be better off under some other set of rules than those enforced in a community. This emphasises the process view, where not only the goodness of exchange or agreement *per se* is a potential source of assessment, but also that it comprises the unintended consequences that can be observed only afterwards.

It should be noted that the opportunity cost explanation discussed here is not applicable to externalities that third parties may experience as a third party cannot withdraw from something she is not engaged in. The basic economic argument is that if third parties are negatively affected, they can always internalise the effect by offering more attractive terms (Buchanan 1975, 38). This assumption is based on the Coasean model of zero transaction costs and full knowledge but it disregards the fact that the way the initial rights are defined affect the parties, however. A third party is imposed costs if the right to create externalities is initially endowed to the exchanging parties. Since the easiness of withdrawal cannot be used as a relevant criterion in cases of third party externalities, it seems that the procedural criterion is the only relevant way to assess goodness. And, as was stated earlier, nothing determinate, independent of the rules that are enforced in a community can be said about such a criterion.

# 4 Conclusions

The contractarian criterion of goodness is essentially procedural. Its normative impact is limited to the examination of the degree of consent to the processes by which rules are established and changed. The argument of the present study that individuals' interests are not limited to consequential considerations but include a procedural justification as well implies that there is a small contractarian within all of us.

The above discussion aims to suggest that we perhaps cannot import explanatory elements independent of their normative connection to the rules of the community. This is to say that there is no cure against infinite regression in a structural analysis of rules. However, it seems to me that infinite regression is not so much of a problem. On the contrary, it actually helps us to understand that there are no perfect solutions, independent of the rules that are adhered to, in social interaction. By introducing an objectivist notion of opportunity costs, contractarian analysis goes beyond the subjectivist, normative individualist principles which limit the assessment of goodness to observed agreement alone. This is an expected outcome since observed exchange alone provides little information to evaluate the whole process starting from the rights, via the exchange itself, to general outcomes. The contractarian pursuit to somehow deal with the assessment of the whole process implies toward a common ground with the evolutionary approach.

Insofar as conventions can be contractual in the Humean sense (Ch 3), and social contracts can be implicit, there seems to be no clear way to define the boundary between convention and social contract, or the hierarchical dominance of spontaneous or designed rules. In the initial state of anarchy this egg-or-hen -type question can be conceptually examined. The view of the present study is that since agreement presupposes mutual expectations, the infinite regression of agreements based on expectations created by earlier agreements can be broken by allowing the initial agreement to be established on expectations that have developed as an unintended consequence of actions taken.

# Chapter5

# Interplay Between Rules and Outcomes

# 1 Introduction

This chapter examines differences and relations between evolutionary and contractarian perspectives. It will be examined to what extent the critique directed toward Hayek's cultural evolution could be counterbalanced by reinterpreting the invisible-hand and group selection arguments of the evolutionary position. I shall suggest that a rejection of spontaneous cultural evolution based on arguments about the instability of PD rules is not necessarily justifiable by the normative individualistic grounds.

The theme of this study refers to a social reality in which aggregate social phenomena cannot be evaluated by reference to outcomes alone. Insofar as the individuals' interests are also directed to justifying behaviour by the degree of conformity to prevailing rules, an assessment based solely on the outcomes remains incomplete. This directs attention to the processes by which rules are established that bring about outcomes. Both the contractarian and the evolutionary approaches recognise the procedural justification of rule making, albeit from slightly different angles.

When separate actions of individuals are examined, the present study considers rule following and case-by-case adjustment as complementary means for the individual in pursuing what she sees proper. The hierarchical structure of rules implies, though, that the individual's choice is never unaffected by some rules. Precommitment to rule following implies that the individual is often motivated to substitute procedural assessment for the consequential one. On the other hand, the change of rules would become difficult to explain by reference to rational behaviour if the individual were not equipped with the capacity to assess expected consequences of separate actions.

The chapter proceeds as follows: section 2 examines Hayek's theory of cultural evolution. Two alternative explanations for the development of appropriate rules are considered: the invisible-hand and the group selection explanations. After Vanberg's critique of Hayek's position, the section will examine possibilities to re-evaluate Hayek's approach. Section 3 examines consequential and procedural elements in social contract and evolution of rules. The common law process provides a field where both consequential and procedural considerations are of import. Section 4 discusses the connection between purposeful behaviour and unintended consequences at the aggregate level. The view advocated in the study is that even though purposeful design is present in collective endeavour, the overall cultural evolution is purposeless.

# 2 Hayek's theory of cultural evolution: an examination of the invisible-hand and the group selection explanations

Hayek's analyses of rules have been examined by many theorists, and limitations and inconsistencies have been pointed out in several occasions (e.g., Gray 1984, Vanberg 1986b, Kley 1994). Both the invisible-hand and the group selection explanations have been found imperfect. A central problem with these explanations is their tendency to view the survival of rules as the measure of their success.

Advocates of the contractarian tradition hold that there is no reason to assume that the evolution of rules provides more desirable rules than what can be achieved through purposeful design. A central argument in this study is that neither the evolutionary nor the contractarian position can provide an account of priority between rules of spontaneous or designed origin. Cultural evolution comprises intentional design of rules by individuals acting separately and collectively. And also, a social contract depends on the conventions that have spontaneously evolved in a group.

## *2.1 Hayek's model*

When Hayek refers to the beneficial working properties of the market as a spontaneous order, he is considering the capacity of the market process to bring about correspondence among the expectations of different people and to utilise dispersed knowledge of the market participants (1948, 77-91; 1976, 107f.). Such an order is beyond anybody's capacity to design and thus its beneficial working properties depend on the appropriateness of rules that constrain the activities of the market participants, which give rise to the order (1973, 43f.).

Instead of being able to provide an independent account on how appropriate rules are to be defined, Hayek refers to the notion that appropriate rules contribute to a beneficial social order. He draws our attention to the processes by which cultural rules emerge and change and tries to find the justification for appropriate rules in the evolutionary mechanisms that create new variants and select among them. Hayek considers two alternative explanations that contribute to the beneficial development of rules, the first referring to the unintended consequences of the interactions among individuals and the second referring to the unintended consequences of competition between groups of people.

### Invisible-hand explanation

In order to be able to provide a convincing invisible-hand explanation of the rules upon which a spontaneous social order is based, one needs to show how the behavioural regularities can themselves be explained as unintended, systematic consequences of a process of interaction among individuals (Vanberg 1986b). A theory of cultural evolution based on the invisible-hand explanation would then need to be able to specify a process of aggregation 'which takes as "input" the dispersed actions of the participating individuals and produces as "output" the overall social pattern' (Ullmann-Margalit 1978, 270) that is to be explained, namely, the rules in question.

To be able to specify a process of rule change as being evolutionary at least two interacting processes need to be determined, namely, the process of variation which continuously produces new patterns of behaviour, and the process of competitive selection which out of the emerging variants systematically selects those that become behavioural regularities in a community (Hayek 1967, 32). For Hayek, the source of variation are the actions of individuals which deviate from the prevailing patterns of behaviour and by experimenting with new practices: 'the law-breakers, who were to be the path-breakers, certainly did not introduce the new rules because they recognized that they were beneficial to the community, but they simply started some practices advantageous to them which then did prove beneficial to the group in which they prevailed' (Hayek 1979, 161).

In order for this model to function one must assume that the deviations are beneficial for the deviators as well as for the imitators who by imitating the initial rule breaker establish or select the new variant. As claimed by Ullmann-Margalit (1978) and Vanberg (1986b), this model is at variance with rules that are beneficial for the group but run against the immediate advantage of any single individual to adopt or imitate it, generally PD rules.

### Group selection explanation

For Hayek, the group selection explanation is related to the invisible-hand explanation although the selection of appropriate rules works at the group level: 'rules of conduct … have evolved because the groups who practiced them were more successful (1973, 18). Resorting to the group selection argument is perhaps due to a consideration of rules that are followed unconsciously and tacitly (Hayek 1973, 43; 1952). It would be difficult to argue that the selection of such rules depends on the individual's assessment of appropriateness.

To refer to group advantage rather than to the perceived benefit of separate individuals, and to argue that those rules prevail 'which lead to the formation of a more efficient order of the whole group' (Hayek 1978, 9) is

different from the invisible-hand explanation, though. The central problem with Hayek's analysis seems to be that he is unable to specify a process by which the fact that a rule is beneficial to a group can be taken to systematically contribute to the existence and persistence of the rule in question (Vanberg 1986b, 83).

There seem to be two distinguishable ways to examine the beneficial effects a rule may have for a group that correspond with the rule's existence (Vanberg 1986b, 83-4): by assuming either that there is a feedback mechanism based on an assumption that individuals are able to recognise a rule's consequences on the overall order, and thus select between desirable and undesirable outcomes; or, that another feedback mechanism is at work at the group level which operates independently of the individual choices.

The latter alternative is claimed to be inconsistent with the invisible-hand explanation because, even though it may be consistent with the idea that rules arise as unintended outcomes, it rejects the idea that social processes can and should be explained in terms of individual actions. The former feedback mechanism is also viewed to be inconsistent with the invisible-hand explanation because it emphasises the significance of *deliberate design* rather than *unintended emergence* (Vanberg 1986b, 84).

### *2.2 Vanberg's critique*

According to Vanberg, the limitations in Hayek's reasoning do not give sufficient reason to reject group selection without consideration of its potential usefulness in describing cultural evolution (1986b, 85). Vanberg's critique is directed to the free rider problem inherent in PD rules, that is, rules that are advantageous to the group in which they are practiced but appear to be disadvantageous to individual members who adhere to them (p. 87).

The paradox of group selection with respect to PD rules is as follows:

> Though individuals who live in groups in which 'appropriate' rules are practiced are better off compared to individuals that live in groups with 'less appropriate' rules, within the groups those bearing the costs of socially beneficial but self-sacrificing behavior would be relatively worse off than those who free ride, who enjoy the group-advantage without sharing the costs of its production (Vanberg 1986b, 87).

Therefore, even though there might be an inter-group advantage from adhering to appropriate rules, there would still be an intra-group disadvantage for those who practice them compared to those who free ride. Given the incentive to free ride, group beneficial PD rules cannot be expected to prevail without conditions that make it beneficial for the

members to adhere to them. This relates to the dynamics of the PD rules examined in the previous chapter.

Reciprocity becomes central to the assessment of whether PD problems are spontaneously solvable. In a relatively small group of people who engage in an ongoing interaction, the mutual exchange of rewards and punishment, of promises and threats has the effect of enforcing PD rules. Immediate gains of defection may be overcompensated by the fear of future losses. Vanberg sees severe limitations to establishing mutual interests and thus 'the mechanism of reciprocity cannot be expected to generate sufficient incentives for cooperative behaviour generally, but under certain restrictive conditions only' (1986b, 96). As the expectation of future interaction decreases and the size of the group increases, incentives to reciprocate diminish. Organised enforcement will be required in order to make cooperative pattern of behaviour viable (ibid.).

### *2.3 Re-evaluation of Hayek's position*

Vanberg suggests that a theory explaining the emergence of group-beneficial behavioural regularities with respect to PD rules would have to show how the conditions are brought about that makes it advantageous for the individual member to adhere to them (1986b, 88). This section discusses the feasibility of such an explanation.

A central issue for providing an explanation for the stability of PD rules refers to what is meant by the notion 'advantage'. I will argue that if PD rules are assumed to fail in providing advantage for the individual member of a group, an organised enforcement mechanism, based on general agreement, fails for the same reason. This is to say that resorting to a third party (such as government) does not provide a solution to the enforcement of PD rules either. An argument by Block and DiLorenzo illustrates what I am aiming at, and provides the rationale to search for a solution in a hypothetical initial state:

> Constitutional economists try to derive a theory of human and property rights from their constitutional framework and they seek to do so on a consensual basis. But how can people give their consent to a contract before it is clear that they have any rights to do so? Where do these rights come from? How can a person agree to be bound by a constitution if it is this very document which can alone establish his rights? If rights are established only by constitutions, then before their advent individuals have no rights. But if they have no rights, what "right" do they have to participate in the construction of a constitution? (Block and DiLorenzo 2000, 571)

Block and DiLorenzo pose the same structural or hypothetical evolutionary question as is present in this study about how the initial social contract can come about if there are no shared expectations start with. This problem is examined in the present study by considering the requirements of an initial state that has developmental tendency away from anarchy. It appears that the development of a cooperation of some kind is pivotal to this issue. The reason why exchange that benefits all parties at precisely the same moment in time does not qualify as cooperation is that expectations of reciprocity are missing. Exchange can occur in anarchy in the sense that two or more participants may collaborate to overturn a powerful party. But insofar as no *temporal asymmetry* in gains from exchange is involved among the collaborators, no trust relations need emerge. As soon as the coup is pulled through, the collaborators are, at least in principle, in a war against each other. Thus, the pivotal factor that needs to be present if this pattern of war is to have any developmental tendency is the presence of temporally asymmetrical exchange. In other words, the members need to be able to cooperate in a way that requires the parties who know each other and who interact constantly with each other to trust that promises are kept. As soon as asymmetric exchange becomes stabilised, the model has some developmental tendency, but not before.

Block and DiLorenzo criticise the constitutional perspective because, by referring to historical facts, one can argue that social contracts have never been based on voluntary agreement, but on 'usurpation or conquest' (Hume 1987, 473). The first governments 'were necessarily the product of war, and thus implied government by one man alone' (Turgot 1973, 69); or, as Edmund Burke put it, 'all empires have been cemented in blood' and that 'the greatest part of the governments on earth must be concluded tyrannies, impostures, violations of the natural rights of mankind, and worse than the most disorderly anarchies' (1968, 53).

Now, my argument for the stability of PD rules was the following: a government is not a solution to PD problems because if the participants do not respect contracts among themselves they do not respect the enforcement by a government. This is to say that the establishment of a government is not based on the assumption of an increase in punishment (both probability and severity), but on the assumption of the government being a neutral mediator. This reasoning is in line with that of Block and DiLorenzo. But when it comes to the historical evidence, it seems to me that their argument is not entirely waterproof.

A central dynamics in anarchy is that coalitions emerge only to be destroyed as soon as any party perceives a slight advantage. Considering the historical account, referred by Block and DiLorenzo, one can argue that what they are showing evidence of is the anarchy part of social evolution, and not, as they assume, of the emergence of a government. It seems intuitional to assume that a voluntary government cannot emerge because a man with a big gun can always come around and break whatever agreement

the members had. The problem with this view is that it somehow assumes the appearance of the man with a big gun as a *unique* event. The man can thus stabilise whatever he likes and the outcome is coercive, like it or not. What this view lacks is the time dimension. We are not talking about any particular man. Rather, we are talking about the method by which stability is pursued. So let us forget about 'the' man and think about a sequence of historical events where men (and sometimes women) with big guns, or with thick wallets, or with huge political power, come forth and claim the power. This is realistic. But the game is then not about establishing a voluntary government; it is about war.

A central problem with the hypothetical state of nature and especially with a conceptual examination of a process leading away from such a state, is that it *necessarily* gets mixed with historical events. The aim is after all to explain or model real events. A possible reason for Block and DiLorenzo's argument that a voluntary agreement is impossible is because they picture the historical events against the background of voluntariness. The overview changes dramatically if the same historical events are retained but related with a model of war.

The present view is the following: coups are part of the game, but do not provide the core explanation for the emergence of a voluntary government. Voluntariness is a *gradual* thing that develops along with history as the members learn to stabilise cooperation. Whoever happens to be in power when the society on the whole stabilises cooperation is of no interest.

### Reconstructing the invisible-hand explanation

I shall now try to describe a process by which the conditions producing a social order that relies upon PD rules could be taken to emerge in a way that provides advantage to the individual member as well as to the group as a whole.

As was stated in chapter three, we need to start from the Lockean state of nature if the process is to have any developmental aspects at all. As a minimum, the members of a social group have to be able to perceive the advantage of the exchange of promises in the form of bilateral reciprocity. At this starting point of our reconstruction, bilateral reciprocity refers to a social situation where the members are able to develop limited trust relations based on cooperation. The golden rule in such an environment is to trust your neighbour with whom you have an ongoing interaction, and who has earned your trust earlier, but to be suspicious of people you do not know. This state of affairs refers to the description that Vanberg would qualify as one where the members are potentially able to stabilise PD rules. As we know, one of the strengths of game theory is the clarity of simple representations. Figure 5.2 illustrates the state under consideration.

| | | B: C | B: D |
|---|---|---|---|
| A | Cooperate | 3, 3 | 0, 5 |
| | Defect | 5, 0 | 1, 1 |

Figure 5.2: A social state with bilateral reciprocity

Even though defection may be the maximising alternative when dealing with strangers, the parties are willing to replace the short-term payoffs with expectations of long-term benefits when it comes to interaction with those they trust. The bilateral reciprocity environment needs to be the starting point. And, as was argued earlier in this study, so it must be for any model that shows any development away from anarchy. Bilateral reciprocity also includes punishment. If being deceived the participant reciprocates by passing any form of punishment (including violence).

The next step in the evolutionary process is that the bilateral reciprocity game becomes gradually extended into a multilateral reciprocity game. Multilateral reciprocity refers to a social situation where the members have learned the advantage of cooperation through experience and do not require favours to be returned immediately by the person a favour was granted. The members are thus willing to endure short-term asymmetry in the balance between favours given and received. There is a longer-term expectation of balance, though.

The condition for such an extension is as follows. Assume the initial social group where not everyone knows everyone else but where the formation of bilateral reciprocal relations may happen simultaneously and unrelated in various places, and also, any member with bilateral relations may create several other bilateral relations. After the bilateral reciprocity has become stabilised as a convention, that is, the members generally expect cooperative response from those they cooperate with, reciprocity stands a good chance of becoming extended into the multilateral version. This is because building bilateral reciprocal relations already contains the asymmetric element that is present in multilateral reciprocity. When starting a bilateral reciprocity game, one of the parties must move first (if they exchanged favours simultaneously, reciprocity could be mixed up with the basic exchange argument). Thus, the first move in the bilateral reciprocity game is structurally analogous to the first move in the multilateral reciprocity game.

The stabilising process of the multilateral reciprocity is described in the emergence of convention. Consider two strangers in a situation where they perceive that a reciprocal relation would benefit them. Both of them already have numerous bilateral reciprocal relations with other people, so cooperation is by no means a new mode of behaviour for them. The only thing they need to resolve is whether or not to trust the other party. The convention of bilateral reciprocity as a precedent would suggest that multilateral reciprocity would more likely become established than not. This is, of course, a generalisation, but insofar as conventions bear behavioural effects, it is a reasonable generalisation. Thus, through social learning, multilateral reciprocity becomes gradually part of the shared expectations. It becomes a convention. Figure 5.3 represents this state of affairs.

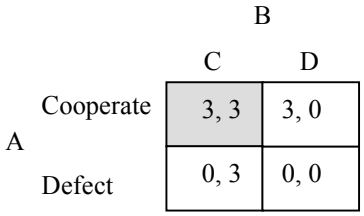


Figure 5.3: Multilateral reciprocity as convention

At this point, the game transforms into a coordination game where cooperation is the expected default response and one would need specific reason to deviate from the common convention. The invisible-hand explanation enters the model in that the individual members aim at reciprocating bilaterally, that is, cooperating with those whom they know. The unintended consequence of a process in which people strengthen and expand their bilateral trust relations lead to the multilateral form of reciprocity. By cooperating, the members continuously recreate shared expectations for future behaviour.

This model relates to Nozick's (1974) analysis of the emergence of protective agency. I argued earlier that a protective agency is not a social phenomenon that can arise in a model of pure anarchy where no rules exist. Therefore, a protective agency does not, as such, solve the problem of instability of PD rules. If a protective agency is viable, then the members already perceive the advantage of some elementary level of cooperation and the possibility for development toward multilateral reciprocity is present.

**Reconstructing the group selection explanation**

Vanberg sees group selection as being inconsistent with methodological individualism and rejects the former (1986b). The group selection idea has been the subject of a number of critiques (cf. Williams 1966; Maynard Smith 1976; Ullmann-Margalit 1978; Trivers 1985). Hodgson (1991) acknowledges the inconsistency between a strict version of methodological individualism and group selection, but instead of rejecting the latter, maintains that methodological individualism should be reinterpreted to permit culture to, in part at least, condition behaviour (p. 78, see also Mayhew 1987). The perspective of this study is closely related to that of Hodgson's.

Hodgson maintains that there is no reason to reject the possibility that selection is operating not only at group level, but at various levels at the same time: among individuals, groups, ownership structures, resource allocation mechanisms, etc. (1991, 79). Thus selection can be based on supraindividual criteria. This perspective complicates two evolutionary ideas:

that cultural evolution is controlled by purposeful design and that it favours efficient configurations.

Consider the above construction of multilateral reciprocity. Assume that there is another group of people who for some reason have not yet developed stable multilateral reciprocal relations and are in the bilateral state represented by figure 5.2. In this state the trust relations are more limited compared to the first group. Cooperation patterns exist, but PD rules requiring spontaneous enforcement by the participants are in their infancy and not relied upon when dealing with strangers.

The first group where multilateral reciprocity relations have become a convention is indeed in a more advantageous position than the second group because multilateral reciprocity as a convention is a *multipurpose* means to cope with all kinds of problems concerning the stability of expectations. Multilateral reciprocity as a convention is not just any rule. It is a central rule by which social life becomes easier and more pleasant. If the members of a society have reasons to trust each other in this multilateral sense, they can discover further improvement in ways that are impossible to attain without trust relations of this type. Thus, the members of the first group will predictably be able to stabilise PD rules as they have learned to prefer conformity to immediate gains from nonconformity. They observe private property because it is in their interests to do so. Due to the rule of private property the members are able to serve each other in a multitude of ways and reinforce expectations of cooperation.

What would be the feedback mechanism in such a construction? It seems rather difficult to argue that the participants are generally able to assess the consequential efficiency of a spontaneous order resulting from private property. It is more realistic to assume, though, that the participants may have procedural interests in observing private property, disregarding its comparative consequential efficiency against some other system. From this perspective, there is no clear feedback mechanism based on consequential issues only. Furthermore, if we permit the operation of multiple selection mechanisms, the picture of cultural evolution becomes increasingly fussy and remote to the idea of purposeful selection.

Assumptions used in the above construction refer to the examination of the rule-guided individual in chapter 2. The individual is taken to be rational, but only limitedly so. Her action is affected by behavioural regularities that are often but not always consistent with rational choice theory (experimental economics in chapter two). She is also taken to value cooperation as a behavioural regularity (Camerer and Knez 1997), and she has grown a sense of fairness (Rawls 1971). She is thus capable of precommitting herself in both the consequentialist sense (Elster 1979) and in the procedural sense suggested in this study. Last but not least, the individual is seen as being able to learn from experience.

One might say that the behavioural assumptions are rather 'rich' in this model. Well, yes they are. Two things are important to notice though:

(1) the assumptions are general behavioural regularities and should therefore qualify as assumptions for a general model, and (2) insofar as by these assumptions the model of interaction changes leading to alternative conclusions, parsimony at the level of assumptions may have prevented something valuable from becoming disclosed.

# 3 Consequential and procedural elements in social contract and evolution of rules

Hayek's approach emphasises the spontaneous, evolutionary element in the processes of rule emergence and change. This view holds 'that the present order of society has largely arisen, not by design, but by the prevalence of the more effective institutions in a process of competition' (1979, 154-5). Vanberg suggests that the resolution to the Prisoner's Dilemma prioritises the *design* of proper rules (1986b). As was earlier explained by Vanberg, a functionalist argument for the appropriateness of rules needs to specify some feedback mechanism by which the members are able to assess the goodness of the general outcomes of rules. He suggests that if such a feedback mechanism functions based on the fact that individuals are able to recognise the beneficial consequences which certain rules have for a group and take action, individually or collectively, to implement and enforce them, then the feedback mechanism cannot be viewed as based on a spontaneous evolution but, instead, on a political process (p. 84). The consideration of the nature of the feedback mechanism directs our attention to the consequences of rules. To what extent the consequential element is *necessary* for the assessment of the goodness of rules becomes a central issue.

I start the examination of the consequential aspect of rules by questioning whether or not the connection between rules and their outcomes necessarily prioritises political process over spontaneous evolution. Consider the general dynamics of a common law process.

**Common law as a spontaneous evolutionary process based on the principle of rule of law**

Common law is often referred to as the judge-made law. This should not be misinterpreted to mean that common law judges make the law as they see proper without reference to the body of law. The proper interpretation refers to the fact that common law making is not a political process of legislation where political and legal experts contemplate the *consequences* of various rules and then impose and enforce alternatives that are *expected* to produce what is aimed at. Instead, a common law process can be seen as emerging in a social group where the group members expect that in cases of conflict, a third, impartial party who acts as a mediator is needed in order to solve the conflict.

The central aspect in the emergence of such a system is not the *design* of the resolution mechanism *per se*, but the *ex ante* acceptance of the principle of rule of law by the members of the group. The acceptance of the principle can be seen as a social contract as the individual members are motivated to accept it due to the fact that, in the face of genuine uncertainty, nobody

wants to deliberately put herself into a situation in which arbitrary coercion is the 'rule' to be used if she ever needs to resort to the resolution by a third party. It should be noted that such an agreement requires expectations of reciprocity. A design of the resolution mechanism does not have behavioural effects if the members are in a war against one another. Another thing that should be noted is that the principle of the rule of law itself cannot be deducted from the agreement that follows the principle. Thus we need to examine the potential processes that can be expected to give rise to the desirability of the principle. Introspection can be seen as the vehicle that carries a type of meta-principle into action. But how that meta-principle is arrived at cannot be explained by introspection itself. Rawls (1971) has suggested that what I have called a meta-principle here is our *sense of justice* that is arrived at by our ability to alienate ourselves from the immediate contextual interests. Although complete alienation may be impossible to attain, an ability to see other people as essentially similar human beings to ourselves may produce a capacity to mentally position ourselves in the place of other people (through the act of introspection). Furthermore, if individual members learn during their upbringing to precommit themselves to a reciprocal pattern of behaviour, a shared sense of justice can be brought about (see further on this issue in chapter two). This suggests that reciprocity is an essential factor without which social contracts, conventions, and the principle of rule of law itself would not have a chance to be established.

A common law process may thus emerge through the expected beneficial characteristics of using a third party as an impartial mediator. The development of a common law process represents spontaneous evolution as well. As a reconstruction of the starting point in a common law process, consider a case where a mediator makes an impartial judgement based on the collectively shared sense of justice. The particular aspects of the cases that follow may be rather dissimilar, but the aim of the mediator is to interpret the sense of justice in separate cases. After a body of precedents has been established they start functioning as guidelines for cases that somehow resemble a particular precedent. A rule to help dealing with numerous partly dissimilar cases may emerge suggesting to 'treat like cases alike' (Barry 1981, 152). Gradually, the normative content of the law that can be seen as being based on the principle of rule of law becomes established.

### Common law and design

Vanberg (1994, 260) suggests that the issue of common law vs. legislation may be related to, but not identical with, the issue of evolution vs. design of rules, since the process by which common law develops, based on decisions made by judges, is different from an invisible-hand process. The main difference assumably is the fact that in the common law process, a judge engages in the purposeful design of rules. Therefore it is difficult to

refer to unintended consequences if there is such a strong element of purposeful deliberation in the making of common law.

Clearly, an interpretation of what is meant by the notion *unintended consequence* becomes central. To illustrate the connection between purposeful, goal oriented action and unintended consequences, consider the market as a creative process (Buchanan and Vanberg 1991) where unintended consequences are pervasive to the extent that the process as a whole cannot be specified as tending towards any predefined goal. This interpretation of the market process is justified even though it is seen as also comprising business firms within which purposeful, deliberate planning and design takes place. As an inference from the market process to the common law process, to say that a common law judge purposefully designs the law is to say that a manager in a business firm purposefully designs the market process.

I shall suggest here that the connection between purposeful action of a common law judge and the unintended consequences that follow can be specified in a way that satisfies the invisible-hand criterion. Although a common law judge *de facto* potentially makes the law in the sense that her ruling may *or may not* become part of the body of common law, the consequences of such rule making are not *essential*, but instead *incidental* (cf. Elster 2000). This is to say that if a particular ruling bears its consequences to other rulings that follow, those consequences are clearly not deliberately designed. The aim of a common law judge is to *interpret* the dispute at hand against the body of law. Although such an interpretation clearly requires purposeful action, the element of purposeful *design* is insignificant. If a judge would want to design a rule that clearly deviates from the existing body of law, such a ruling would be overturned in appeal courts. The notion 'design' in rule making is reserved here to a different type of rule-making process that is discussed in the context of social contract.

The connection between purposeful action and unintended consequences in common law can thus be specified in two distinct ways which both make the interpretation of common law as a process of deliberate design unjustified. First, a ruling is limited to a particular case at hand, the consequences to later rulings of which may remain beyond the epistemic capacity of even a trained judge to foresee. Second, irrespective of the interests of a judge, her task is limited to interpreting the existing body of law.

In a stationary or evenly rotating economy, the need for a continuous flow of interpretation would obviously cease. But insofar as we consider the market process and the common law process open-ended, where the interaction among separate participants produces a continuous flow of novelties and unintended consequences, interpretation and reinterpretation becomes a central feature of learning in both realms (Hayek 1952, 1967; Popper 1972). Discoveries and creations in the market may require novel interpretation in the realm of common law. Equally, as an unintended

consequence, rulings in common law may facilitate new or limit existing ways of interaction and practices among the market participants.

**Procedural interests in common law**

Cultural evolution as a combination of the invisible-hand process and group selection could be reconstructed as follows. A common law process brings about a continuous flow of unintended consequences the goodness of which are not assessed from the consequential perspective but rather through the procedural foundation on which the common law process is based (this makes the justification of a social contract equal with the spontaneous evolution). The individualistic feedback mechanism is therefore not directed to the separate *outcomes* that the common law process brings about, but rather, it is directed to the principle of the rule of law *itself*. The individual members can consider it advantageous for themselves to abide by the rule of law, irrespective of the particular outcomes that the common law process brings about. An understanding of the future as genuinely open-ended motivates the individual to accept the general principle. The individual members may well also be able to perceive the advantage of such a principle regarding the group as a whole. But here again, their interest cannot be directed to the particular outcomes that the common law process brings about. Their cognitive capacity is simply much too limited to permit that.

The group selection part is not unproblematic. The central problem lies in the inability of anyone to foresee how exactly the overall order will be affected by the multitude of particular outcomes that the common law process brings about. This feature can be seen as an analogue to the market process. The market process in its entirety is an abstraction that cannot be expected to be known by anyone. When considering the beneficial working properties of a market order, we need to refer to the rules and processes that bring them about.

At this point Vanberg's second condition for proper justification of a functional argument for the development of appropriate rules comes into play. It refers to the possibility that a feedback mechanism exists that is independent of the intentional action of the individual members, but which may contribute to the advantage of the group as a whole. As Vanberg notes, such a feedback mechanism would be inconsistent with the principle of methodological individualism (1986b, 84). The members may have a procedural interest in conforming to the principle of private property, which then brings about a spontaneous order that is viewed as being advantageous for the group as a whole. The members are allowed to experiment and discover novelties that provide services to the group. It may then also be in the members' consequential interests to retain the convention.

The interpretation adopted here is in line with Vanberg's reasoning. Group selection may well be a real process, that is, that the institutional

framework influences the types of opportunities that are open for the individual members of a group. As an unintended outcome, a particular set of rules applied in a group may bring about beneficial rules, as judged by the members themselves, rather than another set of rules in another group. This type of comparative statement based on the consequential perspective is incomplete, though. Our inability to assess the outcomes in their entirety directs the attention back to procedural assessment. Thus a principle seems to stand out which provides a benchmark, albeit an imperfect one, against which social processes can be assessed. This principle is the correspondence between the preferences of the group members and the rules which bring about the overall order (cf. Buchanan 1977).

### Implications to economic organisations

The dynamics of the model discussed here imply that trust relations are a central issue in economic organisations. An economic organisation can potentially be regarded as the type of social grouping where even the assumptions used in Vanberg's (1986b) analysis suffice to stabilise PD rules. Even large multinational economic organisations can be seen as fulfilling the requirements of an ongoing interaction as it is the interrelations within the *subunits* that influence the behaviour of individual members, not some imaginary abstraction of a relation between the central agent and the tens of thousands of employees.

It is predictable that an economic organisation where multilateral reciprocity has become a convention stands a good chance of producing a constitutional environment that corresponds with the interests of the members. The problem of testing this conjecture is not freed from the dilemma that other factors influence the success of an economic organisation. So success *per se* cannot readily differentiate between good and bad business, as judged by the members. We can easily imagine a profitable firm whose constitutional environment is represented by coercive rules and arbitrary managerial discretion. Thus, multilateral reciprocity is only a principle that facilitates further development. It does not guarantee the goodness of other interrelated processes that deal with, for instance, the quality of knowledge, its complementarities, and its dissemination issues in an organisation.

# 4 Purposeful selection and spontaneous order

Vanberg has raised an important question about how to properly model the 'interplay between "blind" evolutionary forces and deliberate human design' (1996, 690) in the study of institutional development. His analysis is based on the recognition that our theoretical understanding of the market process can contribute to the understanding of how rules emerge and change. As Vanberg explains, Hayek was for some reason unable to extend his analysis of the market process to the realm or rules regarding what is meant by 'appropriate' rules. Constitutional economics has contributed to this problem by suggesting that the appropriateness of rules is to be assessed by observed agreement among the individuals involved.

The point I want to stress in this section is that various interpretations concerning purposeful human action and evolutionary forces are available. My aim is to examine how the concepts of purposeful action, design, and unintended consequences are related.

Commons (1924) suggested that it would be more appropriate to speak of *purposeful* instead of *artificial* selection in cultural evolution. This for the obvious reason that human action is purposeful (see, Mises 1966 [1949]). Similarly, Vanberg suggests that '[t]he market process clearly does not blindly stumble upon problem-solutions that happen to occur in a stream of random trials subjected to some "objective" selection mechanism' (1996, 691). He quite correctly concludes that the evolutionary nature of the market process does not hinge on the 'blindness' of the experimental trials themselves but on its open-endedness toward new variation and competitive selection among variants (ibid.). As Hayek noted, despite the purposefulness of human action, the market process itself is 'blind' in the sense that it is always a voyage of exploration into the unknown (1948, 101). From these considerations, Vanberg suggests that the market process can be specified as a combination of both human purposeful design and evolutionary exploration (1996, 692).

There seems to be room for interpretation with the notions of purposeful *action* and purposeful *design*. It may be that without a conceptual demarcation between these two notions, the latter blends into the former and may bring confusion in the use of the notion design in the realm of rules. The notion of design in the context of the market process does not add anything to the idea that the participants make trials which in the course of the market process are proven successful or less successful. This is because purposeful action *per se* provides a model where individuals have goals which they try to realise by planning and executing plans (Mises 1966 [1949]). To say that individuals design trials is simply to say that individuals make plans and execute them.

Another source of confusion arises from mixing the individual level with the aggregate level. In the market process, all activities take place at the

individual level. To say that the market response to a trial was good or bad is to say that other market participants' responses were such. The market process as such does not respond to anything. The market process is an unintended consequence of interactions among the participants. It is not always clear whether with the notion market process it is referred to the interaction among participants *per se*, or to an aggregate, abstract process that arises from that interaction. When Kirzner (1985) argues that the market process has a tendency toward equilibrium due to the coordinative actions of entrepreneurs, the market process bears an aggregate meaning. Also, when Buchanan and Vanberg (1991) demonstrate that Kirzner's teleological view is unfounded, they obviously refer to the market process as an aggregate level phenomenon. Seen in this way, the market process itself is purposeless. Therefore, it seems erroneous to specify the market process as being a *combination* of both human purposeful design and evolutionary exploration because, first of all, the market process emerges as an unintended aggregate outcome of individual exploration, and secondly, purposeful design is inherent in individual exploration which gives rise to the notion of evolutionary exploration. It should be noted that evolution does not explore anything, only individuals do. There seems thus to be no clear distinction between what is meant by individual exploration and evolutionary exploration. It seems entirely justified to specify the market selection as 'blind' in the sense that the market process itself is 'indifferent' towards the alternative trials that become selected.

**Rules and design**

When we turn to apply the above theoretical understanding of the market process to the realm of rules that constrain and guide actions of the market participants, the notion 'design' becomes especially interesting. I argued above that regarding the market process the term design does not add anything into the concept of purposeful human action. But when we examine the development of rules, the term design receives a meaning quite different from purposefulness. As was discussed in chapter four, rules arising from a social contract are specified as being of designed origin in comparison with rules that emerge spontaneously. Thus to agree upon a rule is to design it. I have been slightly uncomfortable with the notion design even within constitutional realm due to two reasons. First of all, a voluntary agreement upon a rule presupposes a shared understanding of fairness. This is because insofar as the rule in question is to be applied to all members equally, any proposal that would violate the rules of justice in that group would not be agreed upon. And rules of justice may well be of spontaneous origin. The strict unanimity criterion of contractarian philosophy may limit agreement within or within a close range from those rules that are already spontaneously established as conventions. This would then imply that the design element plays a minor role even in social contract.

The second issue has to do with the interpretation of social contract as a discovery process in the sense that any agreement is viewed as being beneficial only hypothetically and therefore open for future criticism (Vanberg 1986). In this case, the term design starts to resemble the basic notion of purposefulness as ignorance about the unintended consequences is introduced, and thus the assessment of goodness is left for the future to disclose. It seems to me that in order for the notion of design to bear distinct meaning from purposefulness, something more is needed. This more could perhaps be found in the type of rule making where the purpose of a new provision is to provide particular predetermined outcomes. In such a case, the rule needs to be designed in perhaps a modular way where the expected outcomes of each module (clause or sentence) are assessed and the unintended interaction among modules is eliminated. The rule itself becomes a designed outcome based on the expected consequences in the sense that the direction of inference is from the expected outcomes to the particular form of sentence or clause. We know the effect that we are aiming at. We only need to articulate it as a formal rule.

# 5 Conclusion

A theme of this study relates to a view that emphasises the human inability to assess aggregate outcomes of social processes in their totality. This directs attention to the procedural assessment that is based on the correspondence between the shared values of the group members and the rules which bring about overall order. Even though the participants aim at rules that would provide beneficial consequences, a central problem remains that the overall order is so complex that a rule's influence to the overall order remains uncertain (Hayek 1978).

When examining the interplay between actions and outcomes, and between rules and outcomes, everything seems to be connected with everything else. Actions are always based on some rules; rules emerge either through intentional design or as unintended consequences of interaction; rules and actions give rise to unintended consequences, etc. The choice of a starting point in analysing such a rule system in its entirety is open to alternative perspectives. What I have tried to argue here is that the constitutional starting point in the structural analysis is not the only justifiable alternative. Starting from an agreement leaves open questions about how the participants recognise mutual benefit in the first place, and why it dominates immediate interests to nonconformity. Analysing PD rules can contribute to our understanding of the dynamics of agreement that is not limited to the notion of exchange.

# Chapter 6

# Rules in Economic Organisations

# 1 Introduction

> Propositions about organizations are statements about human behavior, and imbedded in every such proposition, explicitly or implicitly, is a set of assumptions as to what properties of human beings have to be taken into account to explain their behavior in organizations (March and Simon 1958, 6).

This chapter looks at how organisational rules are interpreted in the heterodox economic literature. The discussion does not necessarily revolve around neoclassical approaches to the firm, such as the New Institutional Economics, because rules in such approaches are normally taken as given by assumption. To test whether the constitutional perspective can contribute to our understanding of the processes by which organisation members come to agree upon certain basic rules of the game, it seems to me to be more helpful to connect the discussion to approaches that share the basic assumptions.

A central finding of this chapter will be that although organisational rules have, to some extent, been analysed in heterodox economics, the literature is largely silent about issues that are emphasised in constitutional economics. These findings imply that there is at least a good starting position for constitutional economics to add value to our understanding of how rules influence organisational behaviour and especially to our perception of the principles and processes by which rules change. The present study as a whole tries to contribute to the assessment of how much value added the constitutional approach can provide for these issues.

This chapter draws upon the contributions of Cyert and March (1963), March and Simon (1958), Nelson and Winter (1982), and Leibenstein (1987). My aim here is not to provide an exhaustive literature review but, instead, to discuss some issues that seem important and upon which other contributions not discussed here can be reflected.

The chapter proceeds as follows: section 2 will view the firm as a behavioural unit. Cyert and March (1963) examine organisational decision-making rules. Among other issues, the findings refer to organisational dynamics where procedural interests dominate consequential considerations. Section 3 views the economic organisation as a balancing act between its own survival and compensation to its members (March and Simon 1958). Section 4 discusses organisational rules as routines (Nelson and Winter 1982). In section 5 Leibenstein (1987) examines the PD and coordination aspects of organisational rules. His perspective relates to the contractarian approach even though his analysis is mainly about decision-making rules rather than about rules of participation and distribution. Section 6 discusses some efficiency aspects of organisational rules.

The central finding in this chapter is that although rules have been, to some extent, analysed in organisational literature, the main targets have been decision-making rules. Since the constitutional rules of an organisation defining the participants' rights in participatory and distributional issues bear their influence as to how decision-making rules are established, a constitutional analysis should receive a central position in organisational study.

# 2 The behavioural theory of the firm

The behavioural theory of the firm, as outlined by Cyert and March (1963), revolves around organisational decision-making processes. The behaviour of a firm is essentially about how to arrive at good choices in an uncertain environment. Organisational choice is heavily conditioned by standard operating procedures, that is, by the framework of rules within which choices are made (p. 99). An organisation behaves much like the individual: it has goals which it tries to attain by available means impeded by imperfect foresight and limited reason of the members.

A major difference between the firm and the individual is that a firm comprises not one mind but several. Whenever we are dealing with collective decision making, two important issues need to be addressed: (1) individuals' *interests* vary both temporally and interindividually, and (2) their *theories* about states of affairs vary in the same manner. This brings increased complexity to the organisational decision-making processes. It seems reasonable that a behavioural theory of the firm needs to somehow deal with the processes by which separate theories and interests are brought together to the extent that we can speak of the behaviour of a firm in the first place.

Cyert and March define their approach as based on the conception that the firm is the basic unit of analysis. The aim is then to predict a firm's behaviour with respect to decisions about price, output, and resource allocation. Their approach emphasises the actual processes of organisational decision making. (p. 19) Before turning to examining these processes, the reader may consider at this point already that before an analysis can begin about how organisational decisions about price, output and resource allocation are brought about, it may be informative to deal with the question about whose theories and interests are to be considered as relevant within the firm. The constitutional approach to the firm suggests that three central rights need to be considered: (1) How the membership in an organisation is defined, (2) which members are empowered with what decision-making rights, and (3) how the outcome of the collective endeavour is to be distributed among the organisation members (Vanberg 1992). Insofar as the main task refers to 'determining the major attributes of decision making by business firms' (Cyert and March 1963, 19), it may turn out that constitutional aspects of the firm play a more central role than is recognised.

### An individualist process approach to the firm

The approach of Cyert and March to the firm is individualist. It is not some single, universal, organisational goal (such as profit maximisation) that they consider when the goal of a firm is discussed. Instead, they emphasise the process approach in that it is the processes by which organisational objectives are defined and changed that become central to inquiry. (p. 19-20)

Insofar as this is the case, viewing the firm as the decision maker may become problematic. The problem I am considering here does not refer to the methodological tension between the firm as an actor and the members who themselves are individual actors as well. The logic of the problem considered here goes something like this: it is justified to consider the firm as an actor insofar as the firm is seen to coordinate actions of its members to the extent that we can refer to concerted action (Vanberg 1992). One can then ignore the internal processes by which goals and decisions are arrived at and consider the firm as the behavioural unit. But if we aim to show how organisational goals are defined within the organisation by the interaction among its members, it may become more difficult to disregard the rights that define how the interaction is established in the first place. To put it plainly, if the standard operating procedures of a firm influence the choices that are arrived at (as in Cyert and March 1963, 99), think about what behavioural influence rules defining who has the right to decide on what and who is to gain and how much will have on organisational decision making and goal setting.

**Organisational goals**

Cyert and March's (1963) examination of organisational goals is consistently individualistic. It begins with a statement that individuals have goals, collectives of people do not (p. 30). The individualistic perspective puts limits to how organisational goals are defined. A reference to collective, mutual agreement is an available alternative but the central problem with this is the fact that people can agree only on rather general goals. This is because as the degree of generality of goals (or rules) decreases the individual members are increasingly able to see how they are personally affected by an agreement upon a certain goal (Buchanan and Tullock 1962). Since interests among members vary conflicts are bound to arise.

Cyert and March suggest three major ways by which organisational goals are reached: bargaining over side payments (political process), enforcing agreements based on bargaining results, and revision of goals by experience (p. 33-41). As will be established in chapter 7 these issues relate to what can be considered the constitutional dynamics of the firm. The process by which the members' interests and theories about alternative goals are mediated and coordinated relates to the process by which a mutual, collective agreement is achieved through compromise and establishment of objective knowledge[9].

The central difference between Cyert and March's analysis and the present one is that their examination presupposes a constitutional framework of rules (participation, allocation of decision-making rights, allocation of rights to the organisational outcome) whereas the present study

[9] Objective knowledge in the sense of shared theories about states of affairs (see Popper 1979 [1972]).

concentrates in examining how constitutional rules are arrived at in the first place.

### Organisational expectations

Cyert and March's (1963, ch. 4) analysis of organisational expectations is based on assumptions of limited reason and imperfect foresight. The discussion on four case studies of organisational expectations implies that consequential efficiency is not a primary goal and that decision making is strongly influenced by non-consequential issues.

In the first case, the management team of a branch plant of a heavy manufacturing corporation aimed at achieving a better safety record. Eliminating fatal accidents and reducing the frequency of lost-time injuries were at the top of the list. After a fatal accident, the management appointed a special committee to assess improvements. The discussion around an important improvement, requiring a considerable investment, shows the ambiguity of organisational decision making (p. 56ff.). The project addressed such organisational dynamics as: the promotion of already favoured projects, lack of information gathering, lack of formally expressed figures (of monetary values), over-optimistic planning, decision making unrelated to costs and returns, lack of assessment of alternative choice options, etc.

The second case is about selecting new premises for a department of a medium-sized construction firm (p. 64ff.). The motivation to search for new premises seemed not been based on considerations of improvements in business processes. Even though centralised operations were believed to be most efficient, the head of the department in question felt that by moving into separate, more remote facilities, some conflicts between his department and others that had been growing could be alleviated. The decision-making process showed that search activity represented a response to some specific events rather than consistent planning. The prospective sites were rejected for various reasons and a systematic comparison of the criteria used in separate decisions was not pursued. And, as the causes for the internal conflicts were largely abolished by renegotiations of earnings contracts of the department managers, the search for new facilities ceased.

The third case is about selecting a consulting firm for a medium-size manufacturing concern (p. 71ff.). At first the alternative Alpha was considered. Alpha being a fairly new consulting firm the managers decided to continue the search, which meant that an offer was requested from an established consultancy Beta. The decision to limit the search to one known company only was made by the controller. Objective comparative assessment of the two alternatives proved difficult. Mutual expectations among the staff and the managers seemed to play a central role here, though. By asking 'the boys [i.e., staff members] to set down pros and cons' of the alternatives, the controller expected to get a balanced judgement by his team members. But what the controller may not have realised was that

by continuing the search after the initial assessment of Alpha, he signalled to his team members that Alpha might not be, for some reason, satisfactory. Thus, the staff members' choice to favour Beta was expectedly influenced by assumptions of the attitudes by the management (p. 75).

The fourth case is related to the third in that the search for a consulting firm was based on a need for examining potential improvements in the company's accounting and merchandising procedures. The fourth case is about the decision making process concerning such improvements (p. 76ff.). Beta consulting suggested three alternatives, each with a different technological solution. A noticeable issue in the process was that in the first round of comparative assessment, two of the alternatives dominated the last one equally. By changing the variables in the second round of assessment, the alternative with centralised electronic data processing became the dominating solution. Cyert and March point out that this was not due to purposeful manipulation of data. Rather, it was due to uncertainty in deciding which costs and savings (which themselves were uncertain) should be counted (p. 79).

With respect to resource allocation, prior commitment rather than marginal return played a central role in the assessment of alternatives (p. 94). Generally, a choice option was accepted once it satisfied the general cost and return constraints and enjoyed the support of key managers, which in turn was based on a complex mixture of personal, suborganisational, and general organisational goals (p. 94-5).

The findings of the four cases suggest that organisational expectations are generally influenced by hope, wishes, internal bargaining needs of subunits, conscious as well as unconscious manipulation of information and expectations, and other influencing behaviour (p. 97). It is predictable that claims for the consequential efficiency of alternatives are used in the bargaining game producing a rather realistic picture of organisational decision making where *ex ante* claims for consequential efficiency are taken as *ex post* evidence of a consequentially efficient organisation. The four cases show, however, that procedural elements (in the form of mutual expectations) are an integral part in organisational decision making.

### Organisational choice

Organisational choices are constrained by the organisational standard operating procedures (aside from the whole constitutional framework). These procedures in turn are seen as reflecting organisational learning processes by which the firm adapts to its environment (Cyert and March 1963, 99). An organisational choice can be decomposed into a decision process constituted by nine distinct steps: (1) Forecast competitors' behaviour, (2) forecast demand, (3) estimate costs, (4) specify objectives, (5) evaluate plan, (6) re-examine costs, (7) re-examine demand, (8) re-examine objectives, and (9) select alternative (p. 100-2). Since all but the last one of

these steps are temporally prior to the organisational choice there remains ample room at each step for influencing and persuasion. A major problem in organisational choices is that the value of internal processes is difficult to measure. I have experience of a large financial institution that experimented with the idea of an internal market. Each department that was engaged in production assessed the value of their respective produce and the consequence was, as is reasonable to assume, an exponentially increasing cost curve as each department wanted to show higher 'profit' compared to the department from which they 'bought' the intermediate product.

# 3 Theory and interest components in organisational decision making

Organisational equilibrium theory (Barnard 1938, Simon 1947) provides the conditions of survival of an organisation. Equilibrium reflects the organisation's success in arranging remuneration to the participants sufficient enough to motivate their continued participation, and thus contributes to the survival of the organisation. The central postulates of the theory are (March and Simon 1958, 84):

1. An organization is a system of interrelated social behaviors of a number of persons whom we shall call the participants in the organization.
2. Each participant and each group of participants receives from the organization inducements in return for which he makes to the organization contributions.
3. Each participant will continue his participation in an organization only so long as the inducements offered him are as great or greater (measured in terms of his values and in terms of the alternatives open to him) than the contributions he is asked to make.
4. The contributions provided by the various groups of participants are the source from which the organization manufactures the inducements offered to participants.
5. Hence, an organization is 'solvent' – and will continue in existence – only so long as the contributions are sufficient to provide inducements in large enough measure to draw forth these contributions.

The motivation for the individual member's participation (postulate 3 above) is essentially the same as is found in the constitutional approach to the firm. The constitutional approach emphasises the voluntary exchange of commitment to the organisation's constitutional rules.

This motivational generalisation is taken here as being acceptable (irrespective of its tautological tendency). The present study will, however, maintain that the notion of voluntariness does not provide an unspeculative interpretation of efficiency. The speculativeness of voluntary exchange arises insofar as the rules that govern exchange can be coercive in the sense that they have been established by a process that bring about Pareto-disimprovement. For instance, an individual may be willing to comply with a certain configuration of property rights if the alternative is ostracism. We do not conventionally refer to Pareto-improvement in a situation where you are forced to buy back your wallet from the thief who stole it. What I am trying to imply here is that although property rights may be fair in the sense that they apply to everyone in a group, they are not positive. Shared values

function as the normative element in any social rule. And what is shared does not have to be shared unanimously. There remains ample room for Pareto-disimprovements in real life.

March and Simon's (1958) analysis on decision making emphasises motivational aspects. They suggest, among other things, that habituation to a particular job or organisation lowers the propensity to search for alternative work opportunities (p. 105). This may or may not be the case. Insofar as habituation releases mental capacity to pursue other things than the task at hand, it may function as the facilitator of discoveries. Also, the present study argues that the possibility to switch between habituation and situational discretion complicates things. Observing a routine, habitual behavioural pattern the observer may be eager to assume that it is the efficient response to environmental factors. And suddenly the pattern breaks as, for one reason or another, the actor's behaviour can better be described as discretion. The observer may again be tempted to put an 'efficient' label to the change of pattern. After all, it was an observable phenomenon and if there is no reason to assume otherwise, the change was for the better.

**Resolution to conflict of interest**

According to March and Simon (1958) conflicts of interest arise in an organisation in two basic ways: due to cognitive limitations and due to dissimilar interests among participants (p. 129ff). The first alternative emphasises intraindividual inconsistencies while the latter deals mostly with interindividual or intergroup discrepancies. Just as was explained in the social contracting process (Ch 4), the failure to agree upon common issues may be due to the fact that the participants simply do not understand what other participants mean; or alternatively, they may well understand but disagree upon the goals that some of the others are striving for.

March and Simon (1958, 129–30) offer four types of organisational resolutions or reactions to these problems: (1) *Problem-solving* assumes that the conflict has arisen due to cognitive discrepancies, (2) *persuasion* can be used to align interests in cases where the goals (interests) of the parties differ, (3) *bargaining* and (4) *politics* refer to situations where the goals differ but where convergence of interests need not occur.

What is interesting from the perspective of the present study is that there is no reference to *rules* as a resolution mechanism to conflict of interest in March and Simon's analysis. The aim of March and Simon's analysis is to emphasise the non-mechanical picture of the organisation member. Such an actor is viewed as being largely autonomous, having goals and preferences different from those of the organisation. Viewing the member as precommitted to organisational rules would then work against the autonomy of the member.

# 4 Organisational rules as routines

In Nelson and Winter (1982), routines refer to decision rules, such as rules of what to produce, procedures of hiring and firing, ordering new inventory, increasing production, procedures for investments and R&D activities, etc. Routine is synonymous to rule insofar as it refers to 'all regular and predictable behavioral patterns of firms' (p. 14). Nelson and Winter's approach to organisational routines is related to that in Cyert and March (1963) in that firms are not viewed as having stable and finely graded means to compare available choice options. Maximising behaviour is thus not assumed in their analysis (p. 36).

Even though Nelson and Winter's contribution focuses on the industry level, viewing organisational routines as genes within populations of firms, they do analyse the dynamics of routines within the firm as well. Rejecting the orthodox view of organisational behaviour as the optimal choice from a sharply defined set of capabilities, they pursue to examine a more realistic framework for choice behaviour in organisations.

Much like the term 'rule' in this study, Nelson and Winter use the term 'routine' in a flexible way. It may refer to a repetitive pattern of behaviour in an entire organisation, or it may refer to an individual skill, or it may even be viewed as an adjective to describe the smooth uneventful effectiveness of organisational or individual performance (p. 97).

**Routine as organisational memory**

Nelson and Winter (1982) view routines as organisational memory (p. 99ff). This is not a self-evident conclusion. After all, we normally think of memory as something residing in our minds, or stored in external memory devices, such as computer hard drives.

The reason for considering routines as a locus of memory derives from a view of organisational knowledge as largely tacit. Organisations remember by doing (p. 99). When an organisation member 'knows' how to do a certain task, such as a billing procedure, she may have learned it by first imitating the procedure carried out by others, and then through a routinisation process internalised as to how and in what circumstances it should be carried out. This view emphasises the interpretation of rules and routines as observable, regular patterns of behaviour rather than rules as codified guides for organisational activities. An organisation then becomes an institution, a complex network of routines whose maintenance and development requires continuity by those who perform the tasks.

### Routine as truce

Routines not only balance on the cognitive aspects of organisational behaviour (such as asking to what extent routine behaviour can be seen as being conscious), but also on the motivational aspects (Nelson and Winter 1982, 107ff.). Organisational members normally hold divergent interests, possess dissimilar power and authority, and are entitled to different degrees of discretion, coercion and decision-making power. Asymmetric power relations and everything that comes with them can cause intraorganisational conflicts of interest. Those with less power would like to enjoy more of it, and those with more power would like to enjoy even more of it.

Nelson and Winter view that largely tacit routines can hold conflicting interests at bay to the extent that cooperative patterns do not break down. The use of the term 'truce' implies, however, that there are some genuinely conflicting interests developing underneath the surface, and that the prevailing truce is of fragile nature (p. 111).

The present study permits the possibility that the dynamics that Nelson and Winter describe as a truce are a more permanent and stable pattern of behaviour among the organisation members – thus qualifying as on-going peace if you will. It is, of course, an empirical matter to what extent organisation members would exploit each other if the circumstances were favourable.

### Routine as gene

The evolutionary perspective advocated by Nelson and Winter emphasises a rather different picture of organisational choice behaviour than what we are normally used to in economics. Whereas analyses at the individual (human, organisation, etc.) level promote an understanding of the actor as essentially free to choose among whatever options she can imagine, Nelson and Winter argue that 'it is quite inappropriate to conceive of firm behaviour in terms of deliberate choice from a broad menu of alternatives that some external observer considers to be "available" opportunities for the organization' (1982, 134). The menu is narrow and idiosyncratic, as well as dependent on the particular routines a firm adheres to.

What the evolutionary perspective does is it directs attention to the view that although organisations (that is, their members) are conceptually free to choose among a myriad of imaginable alternatives, that is not what describes their choice behaviour very well in reality. Routines are path-dependent in the sense that today's modifications are heavily limited by yesterday's routine responses. An analogue from biology would be that although mutations occur, they do not occur completely at random. An elephant does not grow wings, assuming natural selection or not.

# 5 Prisoner's Dilemma and convention in organisational decision making

Leibenstein's (1987) analysis of organisational decision making bears close resemblance to the main theme of the present study: Prisoner's Dilemmas (PDs) are an important part of collective action that needs to be taken into account, and conventions solve, to some extent, PDs in the sense that they prevent the disadvantageous PD dynamics from arising.

Even though Leibenstein uses the same tools as is used in the present study, that is, mixed-motives and coordination games, his analysis differs from that of the present study. Leibenstein examines organisational decision making regarding the members' choices whether or not to put effort either as a manager or as an employee, whereas the focus of the present study is on the dynamics of rule emergence and change at the constitutional level. Therefore, the source of stability of conventions and the resolution to PDs are discussed in a slightly different tone.

### Tension between the manager and the employee

Leibenstein (1987, Ch 5) analyses organisational conflicts of interest using a clear dichotomy between the manager and the employee. A maximising manager would want to have employees performing as well as they possibly can while paying them as little as possible. A maximising employee would in turn want to perform as little as possible without causing her to be discharged. A fully cooperative pattern would be one where the managers treated the employees as well as they possibly could given the firm's resources, and where the employees performed at the peak of their abilities having their goals in complete harmony with those of the firm's. Leibenstein calls this configuration the 'golden rule'. An intermediate form of cooperation would be one where the employees perform according to some average level of effort and the managers compensate them according to the same average performance level.

Leibenstein concludes that the intermediate form is likely to become stabilised for the following reasons. In a golden-rule configuration, the opportunities for defection are too ample for both sides. The probability of exploitation refers to the Western individualistic culture, which is prone to view many forms of social interaction through the divide between winners and losers (p. 53). Two basic assumptions of the model emphasise the tendency toward a less cooperative pattern as well: both sides make their decision on their future behaviour simultaneously without knowing the other party's choice, and it is assumed that the cooperative pattern breaks immediately if even very few of the members start deviating.

According to Leibenstein, one of the main reasons why the maximising pattern is avoided is due to cultural and intraindividual reasons.

If cheating is generally condemned in a society, the members may not be willing to pursue such behaviour even if opportunities were available. The organisation member may not be willing to 'change gear' and act against her innate values. If individuals 'cannot artificially impose changes in their feelings about cooperation' (p. 54), it is expected that they cannot do that about defection either.

The interplay between the individual's self-interest and cultural pressures in the form of conventions is rather interesting in this context. Even though there may be cultural and intraindividual pressures to cooperate, such pressures are not taken into account in facilitating the golden rule configuration. On the other hand, self-interest is assumed to bear an enormous influence when the stability of a cooperative pattern is considered. It takes only very few observations of defection and the individual is willing to follow suit, and, inconsistently enough, throw all her cultural and intraindividual values into the bin. Yet still, as soon as we approach the noncooperative alternative, it is time for the self-interest to get thrown out of the window and in comes the cultural and intraindividual cavalry.

The use of self-interest is not only speculative for the above reasons. There may be a solution to the above PD that permits the use of self-interest in its strictest form, and without the helping hand of conventions. Interestingly enough the solution requires precisely the Western individualistic, winners-and-losers culture as well. It is realistic to assume that managers do not compete only against employees. They compete, it seems to me, more against each other. Retaining the PD dynamics, a manager who wants to build up her career arguably wants to stand out from the crowd. Insofar as we assume that success as a manager derives from how well she is able to motivate employees to contribute, there should be enough incentive for any individualistic and aspiring manager to treat her subordinates well.

At this point a central controversy in Leibenstein's model becomes apparent. The model is by no means individualistic as it assumes the managers and the employees to behave as completely homogenous members within their groups. What I wrote above about the individualistic manager applies to the individualistic employee as well. Insofar as there is self-interest at play (and there surely is), I would argue that it is more generally directed to building a good reputation and career. For that purpose, self-interest makes employees thrive and compete to stand out in the crowd by making contributions that are recognised as valuable by the management. Retaining the realistic assumption that mostly those contributions are recognised as valuable by the management that contribute to the attainment of the organisational goals, the strict self-interest *per se* can in fact and does provide reasonable incentive for the employees to align their interests with those of the organisation.

As a general observation it seems to me that when shirking and free riding is discussed in models and theories that emphasise their presence in firms, incentives are limited to those of financial nature. If the employee is viewed as being unable to benefit financially from excellent performance (e.g., due to the terms of her employment contract), it is assumed that her maximising response will be to stop aspiring at the 'optimal' level balanced by the compensation level. If this were the case, what is wrong with most employees I know who, of course, expect their compensation to increase in time but not nearly as often and as directly and simplistically as is assumed. This view hinges on the (in)ability of the individual to perceive long-term consequences. I argue that it is the *uncertainty* about when one is able to capitalise on past performance (i.e., reputation) that makes employees willing to delay gratification and perform constantly well. To put it more boldly, it would be against ones self-interest not to perform as well as one can because, assuming the same optimality framework that often underlies incentive discussion, one would then not maximise the discounted rewards being accumulated throughout the entire career.

To what extent my above description of the truly individualistic managers and employees is viewed as realistic depends, of course, on what type of employees and managers we have in mind. And furthermore, what type of business culture is considered relevant. Leibenstein's analysis is based on North American culture founded on a strong tradition of individualism. Generality is a prominent criterion in testing the goodness of a model or a theory. Testing whether American employees in general maximise by shirking or by building reputation and career is certainly difficult, considering the heterogeneity of the group labelled as employees. This difficulty may provide ample room for models in both directions to be considered relevant, and emphasise the usefulness of partial theories.

### Organisational conventions as the remedy for PDs

Leibenstein's (1987, ch. 7) analysis of organisational conventions examines, e.g., effort conventions among the employees, and those of working conditions and wages. These conventions are treated in the same manner that is analysed in chapter 3 of this study. They are enforced spontaneously by peers and may have strong inertia against change (the difficulty of accumulating critical mass when acting against the *status quo* is disadvantageous for any member).

It seems obvious that if, for instance, effort conventions exist compelling all employees to perform at a certain performance level, PDs are avoided. Leibenstein's model shows that individual aspirations toward *higher* or lower levels of performance would be retaliated by peers. Thus the *status quo* remains stable.

Two questions may be interesting in this context: what does it mean to say that conventions are remedies to PDs? And, is it reasonable to assume

that performance conventions are generally stable (or that they exist) in Western firms?

In the present study, conventions are viewed as preventing some PD dynamics from arising. But the working properties of such conventions are not assumed to prevent activities which are essentially competitive. In other words, and drawing on my discussion above, assuming an individualistic culture, it seems to me that there is no reason to assume that peer pressure in *limiting* performance is the key factor that influences the employee's aspiration level generally. This is because if such conventions generally existed, it would make it much easier for any given employee to stand out in the crowd and get promoted than in an environment where such conventions were not present. Such a convention would be a *reverse* PD game. Anyone defecting would not only gain herself but also increase the social benefit. This is to say that a convention of this type does not fulfil the requirement of being obviously against anybody's self-interest to deviate. Leibenstein's conclusion is, however, consistent with his view that 'most people do not wish to appear to "stick out" in their behavior' (p. 82). If that were the case we would experience nearly evenly rotating economies around us, and to formulate it by using the concepts of this study, most organisation members would base their choices on *procedural* considerations alone.

# 6 Some efficiency considerations

Four alternative interpretations of organisational efficiency are considered here:

1. An economic organisation is efficient insofar as the members' goals are consistent with those of the organisation.
2. An economic organisation is efficient insofar as it chooses those alternatives that produce the largest results for the given application of resources.
3. An economic organisation is efficient insofar as it is profitable in the long term.
4. An economic organisation is efficient insofar as it provides Pareto-improvements for the stakeholders.

Theories of bureaucracy, dating from Weber (1946, 1947), analyse the consistency between organisational goals and those of the members'. A general aim in these theories is to analyse the intended as well as the unintended consequences of various organisational control mechanisms.

Merton's (1940) analysis emphasises the dysfunctional aspects of organisational learning. The human ability to infer solutions from past experience to novel situations may cause unanticipated and undesirable outcomes for the organisation. This gives rise to the demand for control, and for reliability and accountability of behaviour within the organisation. The solution comes in the form of standard operating procedures.

Where Merton emphasises rules as a response to the demand for control, Selznick (1948) emphasises the delegation of authority. Delegation seems to both increase and decrease the discrepancy between the goals of the organisation and those of the participants. Through increased training in specialised competences, the members are better able to solve problems relating to their limited areas of expertise. On the other hand, delegation may provide opportunity for subgroups to develop their own goals that do not match with those of the overall organisation.

Gouldner (1954) provides some very interesting insight into the working properties of organisational rules. Like Merton, he is concerned with the consequences of bureaucratic rules for the maintenance of organisational structure. The use of general and impersonal rules regulating work procedures brings about a decrease in the visibility of power relations within the organisation. This affects the legitimacy of the supervisory role of the management and reduces the level of interpersonal tension within the organisation. These rule further the survival of the work group and are therefore reinforced.

Although Gouldner analyses rules that regulate work procedures, his view bears close resemblance to constitutional rules as well. Constitutional rules defining the basic rights among the participants are effective by the

same rationale that Gouldner uses. When the basic rules are in place the members can concentrate on competing within the fields of productive activities. It should perhaps be noted that Gouldner's model does not assume this type of constitutional argument, quite the contrary. His reasoning follows the same line as is discussed in Leibenstein's analysis above. For Gouldner, work rules provide the participants with information about the minimum acceptable behaviour, and that is where the equilibriation process will lead to unless close supervision is not applied.

Bureaucratic models of organisational behaviour generally emphasise the first interpretation of efficiency, that is, the goal congruence version. This does not mean that survival aspects are not addressed at all, though. For instance, Merton's model recognises the problem that rigid rules and authority trappings may have regarding customer satisfaction, and through that to the consequential efficiency of the organisation. A central problem with the goal congruence version of efficiency is the fact that the organisation as such cannot have goals, only its participants can. Thus, rules defining who is entitled to define organisational goals become the central factor in defining the organisational goals.

Simon's (1947) interpretation of efficiency is in line with the second version of efficiency. A central problem that economic as well as non-economic organisations face in decision making is the fact that costs and revenues are often interdependent in ways that make the consequential assessment of alternatives difficult. Should an organisation choose a high cost—high revenue alternative over a low cost—low revenue one? A decision of this kind may employ the third approach to efficiency, that is, considerations of risk of failure and survival.

In Nelson and Winter's (1982) analysis, profitability of routines provides the central criterion of success (p. 121). Their evolutionary perspective does not claim that any existing configuration is globally efficient, however. The firm's 'success and failure depend on the state of the environment. As long as the world rewards great tennis playing, great tennis players will succeed in the world, regardless of their talents as physicists or pianists' (p. 134–5).

The fourth version of efficiency is perhaps the most complicated. Two central factors make an assessment problematic. First, Pareto-efficiency considerations are not limited to monetary flows, and, as will be argued below, not even to survival aspects. Second, the heterogeneity of interests among stakeholders makes it difficult to constantly arrive at Pareto-improvements. The concepts of procedural and consequential interests used in this study may help to illustrate the complexity of reaching Pareto-improvements.

The general interpretation of the Pareto-criterion implies that only those changes in social arrangements are acceptable by which at least one person is made better off without someone else being made worse off. A central question arises about what is meant by making someone better or

worse off. The general way to deal with this question refers to the *outcome* of a choice (consequential consideration). A choice leads to Pareto-improvement insofar as its consequences meet the Pareto-criterion. The problem with this approach is that consequences occur after the choice is made and a choice to include or exclude unintended consequences needs to be made. Insofar as the Pareto-criterion is supposed to refer to *real* increases or decreases in the welfare of the participants, unintended consequences should be included. Otherwise the concept becomes empty of content.

An alternative way to interpret Pareto-efficiency refers to the principle provided by constitutional economics. A collective choice provides Pareto-efficiency insofar as it meets the unanimity criterion in agreement (procedural consideration). If the participants who will be affected by the choice agree upon a certain rule or action, the agreement per se provides the assurance that a Pareto-efficient alternative is reached. It is, however, clear that as much as it is difficult to foresee the unintended consequences of a choice, it may be as difficult to foresee who the relevant participants are that will be affected by a choice.

The procedural approach to Pareto-efficiency *per se* does not differentiate between successful and less successful outcomes. Therefore, a procedurally efficient choice can lead to a complete failure killing the organisation altogether. Thus the above view that Pareto-improvement does not ensure survival.

The strict constitutional criterion is not easily applicable when considering societies in their entireties, but is perhaps more applicable with regard to economic organisations. The limited size of an organisation and the conventions that provide conformity facilitate the establishment of mutual ground among the members. Against this background, it is interesting to see that the constitutional approach has not been applied much in analysing the economic organisation. To be sure, the strict constitutional criterion of goodness is compensated within economic organisations with two alternatives that are based on common sense reasoning but which do not guarantee mutual benefit.

The first alternative refers to the Hicks-Kaldor criterion by which a change in rules would be acceptable insofar as the benefit would over-compensate the loss and those who win could hypothetically compensate those who lose. The contractarian position argues that in such a situation the compensation should be observed, otherwise no guarantee of mutual benefit is provided. The actual compensation would then change the payoffs so that everyone would gain and the model would transform into a Pareto version. In organisational life we can observe the Hicks-Kaldor –type reasoning all around us, though. One alternative is chosen over another based on an argument that it provides superior payoffs. Insofar as such an argument is based on the view of the majority (in the power-using capacity sense), the criterion transforms into the majority rule.

As was argued in chapter 4, according to contractarian reasoning not all choices among rules need to satisfy the strict criterion of unanimity. It is entirely justifiable for any group to unanimously agree upon relaxing the criterion for post-constitutional rules that are specific to the extent that a complete agreement would be too costly to achieve (Buchanan and Tullock 1962). This makes the perspective more operational but also creates a logical problem. If post-constitutional rules may be based on various degrees of unanimity (majority, two-thirds, three-fifths, etc), then a choice of which category to use regarding a particular post-constitutional rule becomes central. If a choice of the category to be used with a certain rule is *less than unanimous*, the choice itself is unjustified on constitutional grounds. All the participants know that the higher the degree of unanimity that is required, the less probable it is for a rule to become accepted. Thus those who favor a certain post-constitutional rule try to get it into a category with a low degree of unanimity, whereas those who oppose it try to get it into the category of the highest level of unanimity. Thus, if such a choice *itself* is not unanimous, there is no guarantee that any post-constitutional rule is efficient. As far as I know this issue has not been discussed, even though it deals with the central logic of constitutional reasoning.

# 7 Conclusions

The contributions discussed in this chapter share a set of basic assumptions about the human agent and the type of interaction that goes on within a group of such agents. Imperfect foresight, limited reason, subjective knowledge, and rule following describe the general qualities and behavioural regularities that the human actor manifests in these analyses. The present study shares these assumptions and tries to analyse the rule-following assumption more thoroughly.

Even though organisational decision making is analysed in the light of decision rules, these contributions are silent about rules that define the basic, constitutional rights of the participants to pursue such decision making in the first place. As stated earlier, constitutional rules define participation in the organisation, the right to decision-making processes, and the allocation of the organisation's outcome. These rights constitute the basic structure and the working properties of an economic organisation. Without their presence we could not perceive something to be an economic organisation.

Since the constitutional rules of an economic organisation influence what kinds of decision-making rules will be established, analysing decision-making rules alone does not provide a satisfactory view of the impact that organisational rules have upon organisational dynamics. For instance, careless (re)design of options schemes in organisations may result in negative unintended consequences as the designers fail to acknowledge that such schemes carry important constitutional impact as well.

As a limitation to the constitutional perspective the chapter argues that it is not permissible to jump from the unanimity criterion into a sub-unanimous alternative and simultaneously retain constitutional justification. Saying that sub-unanimous rules fulfil the constitutional criterion only because at some constitutional point in time the members decided upon a *principle* of post-constitutional rule making is not considered satisfactory here. The principle *per se* does not distinguish between justified and unjustified rule making; what does is the process by which post-constitutional rules are decided. Insofar as a choice of the category is less than unanimous (which is expected to be the case since the consequences are assumed to be more readily assessable with post-constitutional rules), no guarantee of mutual benefit should be expected.

# Chapter 7

# Extending the Constitutional Approach to the Firm by Introducing Conventions

# 1 Introduction

This chapter examines the applicability of an extended constitutional perspective to business firms. The firm is constituted by a group of self-interested people cooperating and competing within a set of multi-layered rules. Individual decisions and actions are interrelated and coordinated in ways that allow us to refer to *corporate* (Coleman 1990) or *concerted* (Vanberg 1992) action. The contribution of the constitutional approach is that it highlights the (explicit or implicit) constitutional agreement as an *exchange of commitments* (Vanberg 1994, 140). The contracting parties benefit from constraining their future choices within the constitutional framework. The core argument of the constitutional perspective to the firm is that an organisational social contract results in relations among the parties that are different in kind from market relations (ibid.).

The constitutional perspective emphasises the choice among rules within which interaction takes place. It directs analytical interest toward the justification of processes through which rules are chosen. The emphasis on the choice among constraints, instead of the choice among alternatives within specified constraints, is what distinguishes constitutional political economy as a research programme from conventional economics (Buchanan 1991). Buchanan illustrates this distinction by referring to games. Games are defined by their rules. Before a game can begin the rules must be decided upon. After the rules have been agreed upon, the game is supposed to be played according to those rules. The constitutional perspective emphasises this two-step notion of constitutional rules. My attempt in this chapter is to extend this view further. The game metaphor does not connect the rules of the game with the system of already existing rules. My project is to try to provide a connection between the initial first step of defining the rules of the game to conventions that may provide further explanation to the type of rules that can be agreed upon. I will try to show that introducing conventions into the realm of constitutional economics is not only consistent but, in fact, can be beneficial in providing explanation as to how the members perceive mutual benefit to start with.

An introduction of conventions into the realm of constitutional economics does not lessen the importance of agreement as the ultimate criterion of goodness. The convention perspective can, however, provide a partial explanation for the emergence and perseverance of heterogeneous organisational constitutions. Organisation-specific conventions provide the interpretation of rights and obligations among the participants. It is through these conventions that the participants perceive the meaning of constitutional rights and obligations.

A conjecture of the chapter is that conventions, e.g., on fair distribution, vary across firms. Also, similar distributional schemes may result in surprisingly different outcomes, depending on what other

conventions are adhered to. In some organisations, a piece-rate incentive system may result in shirking and misrepresentation (Cf. Whyte 1955), while in some others, it may work in the opposite direction providing powerful motivation for the employees. Miller (1992) suggests that these differences may be the result of differing expectations and beliefs among the employees about the extent to which the managers will respect work contracts.

There are not many theorists who have analysed the firm from the contractarian perspective. Vanberg (1992) has provided a persuasive analysis on how the constitutional paradigm can provide a consistent individualist interpretation of organisations as acting units. His approach is linked with Coleman's (1990) analytical perspective on the procedural foundations of collective action. Gifford's (1991) constitutional analysis of the firm argues that the firm will benefit if relation-specific investments can be secured through the owner's attempt to purposefully design an efficient constitution. Wolff (1997) recognises that corporate culture, as in Kreps (1990), can be taken as an implicit part of the constitution of a firm. Langlois (1995) discusses the interplay between constructed and spontaneous elements in the emergence and perseverance of firms.

Figure 1 illustrates the goal of this chapter. The contributions of Coleman, Vanberg and Gifford are depicted as the area A, providing a constitutional approach to the firm. My goal is, first, to justify the introduction of evolutionary themes into the contractarian perspective (area B) and, second, to provide conjectures on the applicability of the extension to the firm (area C).

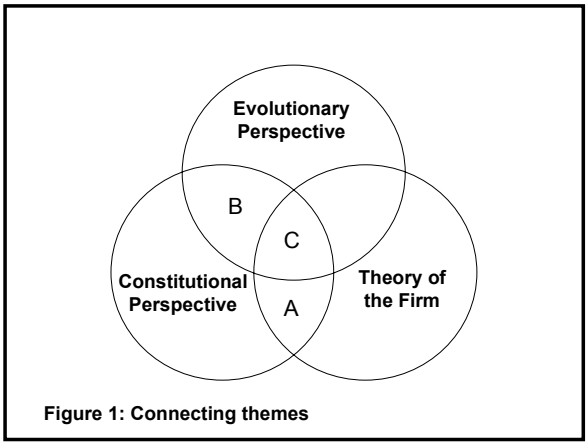


**Figure 1: Connecting themes**

The extended constitutional approach may have some interesting connections with other theories of the firm. I propose that it can be viewed as an overarching perspective that can encompass many existing insights into the theory of the firm. It can provide further explanations on the resolution of *coordination* and *motivation* problems within organisations, emphasised by the modern contract theory (Cf. Milgrom and Roberts 1992). Matters such as the feasibility of incomplete contracting, or why participants are willing to constrain their rent-seeking and opportunistic tendencies in

relations characterised by asset specificity and asymmetric information, receive explanations that complement transaction cost and contract theories. Corporate culture has a close connection to the present theme and will be discussed in connection to constitutional reasoning. The chapter is, however, limited to general discussion and cannot provide a systematic account for all relevant matters because the central contribution of the chapter has to do with extending the contractarian perspective.

This chapter uses the notions of *rules* and *institutions*, *norms* and *conventions* extensively. These notions have different meanings in different approaches so it is perhaps beneficial to try to define their meanings in the present context. All these notions share a common denominator in the concept of rule-following behaviour. One way or another, they all deal with adherence to rules at some level. The term 'rule' is used as a general notion and the context in which it is used will then delineate the particular meaning. Institutions have various interpretations. For Hayek, an institution is a rather abstract system of rules of conduct (1967). Polanyi (1958) emphasises the implicit or tacit nature of institutions. The constitutional perspective of Buchanan and Tullock (1962), and Brennan and Buchanan (1985) emphasise institutions as rationally constructed constraints within which interaction takes place. As Langlois (1992) suggests, these different interpretations do not have to exclude others but can operate within a system of social institutions.

The chapter is organised into five main sections. Section 2 will provide some basic principles of constitutional economics. It will also initiate the rationale for extending the constitutional perspective with conventions. After that I turn to examine the constitutional theory of the firm. The discussion is limited to the contributions that are closely related to the contractarian reasoning provided by Buchanan and other advocates of constitutional political economy. The reason for this limitation is based on the recognition that the term constitution may be used in various ways that do not have to correspond with contractarian philosophy. In section 4 I will examine the extended constitutional perspective further in relation to the theory of the firm. In section 5 I will then discuss some general connections between the present approach and other theories of the firm. The suggested connections are far from being inclusive and many aspects that perhaps should have deserved space may have been left out because of the sheer epistemic limitations of the author. Finally, in section 6, some concluding remarks will be discussed.

# 2 On some basic principles of constitutional economics

Constitutional economics is essentially about the examination and evaluation of the foundational rules of social order. It is an inquiry into the interrelation between what Hayek called the *order of rules* and the *order of actions* (Hayek 1973). The constitutional perspective suggests that in our pursuit for social improvement, changes in the order of rules ought to be the principal means (Vanberg 1994, 5). It directs our analytical attention toward the *choice among constraints* (Buchanan 1991, 5) This perspective implies the recognition that societies are complex systems where purposeful design directed to particular outcomes does not in many cases bring about what is desired. This results from the genuine uncertainty of outcomes. The source of uncertainty is our ignorance of the unintended consequences inherent in human (inter)action. Although an organisation can be viewed as being intentionally constructed to realise a certain purpose, the actions taken within the organisation have unintended consequences as well. This relates to Hayek's view on the relation between the origin of rules and their outcomes as he states that 'it is possible that an order which would still have to be described as spontaneous rests on rules which are entirely the result of deliberate design' (Hayek 1973, 46). What is meant by the notion of *purpose* becomes central. If by purpose it is referred to rather concrete ends, it is consistent to view the goal of an organisation as an outcome of purposeful design. This, however, leaves open the question to what extent the attainment of that goal can be viewed as a planned process. Alternatively, if one views purpose as being directed towards more abstract ends such as self-maintenance or survival (cf. Selznick 1948), the purpose seems to owe more to spontaneous elements.

The constitutional perspective directs the analytical interest from the goal-oriented discussion to the foundations of agreement on participatory and distributional rules. The firm is then not defined through its possible goals, but through the rules that constitute a system of productive relations among the participants. The constitution of an organisation specifies the terms of participation: (1) which resources participants are to contribute to the organisation, (2) how and by whom the decisions on the use of pooled resources are to be made, and (3) how the resulting benefits from the joint endeavour are to be shared among participants (Vanberg 1985, 22).

Constitutional analysis is consistently individualistic. (1) The derivation of institutional constraints is based on a calculus of individual interests. (2) Collective choice is derived from the participatory behaviour of individual members. (3) Emphasis is directed to the selection of rules that will limit the behaviour of those who operate within them (Buchanan 1991, 8).

The concept of normative individualism provides an ontological point of departure for constitutional economics advocated by Buchanan and other

contractarians. Normative individualism suggests that we should take the values and interests of the individuals involved as the relevant standard against which the goodness of rules and their outcomes is to be judged (Vanberg 1994, 1).

The constitutional perspective highlights voluntary exchange as the core motivator for the individual to limit her behaviour within constraints (Buchanan 1991, 5). The cost of limiting one's own behaviour is accepted insofar as it does not exceed the benefit resulted from reciprocal behaviour of others. This perspective emphasises the calculative rationality of the individual who actively chooses her own constraints. By definition, a voluntary exchange happens only when the participants expect to gain from the trade. What constitutional economics does is it brings exchange within the realm of collective decision-making processes.

The subjectivist position of the constitutional perspective recognises that values and theories about the world vary across individuals. This limits efficiency considerations because it is believed that no supraindividual scalar of goodness exists. There is no reason to believe that the ordering of preferences would not vary across individuals.

From the subjectivist position, an assessment of efficiency relies on revealed preferences of the individual. When the idea of voluntary exchange is transferred to the realm of collective choice, the strict criterion of revealed preferences through observed exchange needs to encompass all the parties. As the subjectivist position holds that the values of individuals are incommensurable, an exclusion of any one party from the exchange breaks down the possibility to verify that the observed exchange was in fact efficient.

Since individuals vary in their tastes and interests, it is likely that when a group of people get together in order to pursue something collectively, conflicts of interests arise and mutual agreement may thus be difficult to achieve. The members need to *compromise* before a mutually agreed solution can be reached. The solution may not match perfectly with anybody's personal interests but provides a more desirable outcome than being without it. The question about how to facilitate a compromise thus becomes central. A compromise requires the parties, to some extent, to alienate their own self-interests and, through introspection, assess what would be considered fair by the other parties.

The constitutional bargaining process itself contains aspects that facilitate agreement (Brennan and Buchanan 1985, 29). Rules are by definition more general than the outcomes that result from action guided by those rules. A constitutional choice among alternative rules contains the elements of generality as a chosen rule needs to be applicable in numerous contingencies. Another basic characteristic of a rule is its extended time dimension. A rule needs to be applied over time, otherwise it can hardly be considered a rule. Due to these considerations, the individual faces genuine uncertainty about how her position will be affected by the operation of a

particular rule. Insofar as mutual agreement is the goal, the individual tends to agree on arrangements that can be considered fair in the sense that they are broadly acceptable (ibid., 30).

**Introducing spontaneous elements**

For a contractarian, the only justified criterion of goodness in collective choices is a unanimous agreement among the participants. Alternative constitutional arrangements can be analysed in a hypothetical initial state to discover what basic principles such rules should fulfil. This may help her to create new alternatives that may receive acceptance among the relevant group. But, as Buchanan has recurrently noted, the members of the group are the sovereign decision-makers whose individual values are the only justified source for efficiency considerations. What makes this central principle problematic is its consequence on innovative and creative aspects of social endeavour. It is obvious that many innovations are such that only few understand their potential value immediately. Changes in rules are especially difficult to negotiate because individuals generally value the *status quo* (see, e.g., Schlicht 1998). Another problem that, e.g., Barry (1984) has pointed out, arises as all the members are in a position to veto an alternative that would otherwise be desirable. From the contractarian position, there is of course no such a thing as 'otherwise desirable', but it is intuitive to think that somebody may want to veto whatever the rest of the group are suggesting. This connects to the pragmatic criteria of justifiable exclusion from decision-making (children, mentally challenged, etc.). Any agreement on the contents of such a list fails by necessity to meet the contractarian ideal. This is because the exclusion must occur before the list is agreed upon.

The evolutionary approach, defended by Hayek, appears to be able to resolve this problem of agreement. It is not at all decisive to what extent a new creation happens initially to be accepted in a group. The success of an institutional arrangement is not judged by its emergence, but by its dissemination and survival. Now survival is a stimulating concept in institutional evolution, especially in the writings of Hayek. It connects us to efficiency considerations that deserve some discussion in connection to both these approaches.

Hayek's 'twin concepts of evolution and spontaneous order' (1979, 158) suggest that, (1) the separate inter(actions) among individuals produce an overall order that can be viewed as an unintended evolutionary outcome, and that (2) the abstract rules that facilitate the emergence of this order are themselves the spontaneous outcome of an evolutionary process. The first notion refers to the knowledge problem of society as the overall outcome 'will depend on a very large number of particular facts, far too numerous for us to know in their entirety' (Hayek 1973, 23-4). The second notion suggests that most rules that facilitate the evolutionary process are not of designed origin. This recognises the cumulative nature of knowledge inherent in the

spontaneous rules, 'which are the product of a process of evolution in the course of which much more experience and knowledge has been precipitated in them than any one person can fully know' (Hayek 1967, 92). This argument seems at first sight to contradict the constitutional emphasis on explicit agreement. The contrast is apparent if social contract is viewed as an ahistorical, unique agreement, but if one takes social contract as an *ongoing* process the contradiction seems to diminish. An explicit agreement may be reached precisely because the participants use knowledge inherent in rules, accumulated by the path of history, that they would separately be unable to obtain.

In Hayek's work, there is a tendency to see the spontaneous evolutionary process as superior to intentional design. For Hayek, the selection of rules is 'guided not by reason but by success' (1979, 166). A central problem with an *a priori* statement about the comparative advantage between evolution and design is that it is difficult to establish why experimental activity as an individual pursuit should be any better or worse than a similar trial and error process as a collective endeavour[10]. Insofar as we see social contract as an ongoing process, a group of people may, through team effort, reach a more refined rule than would be the case by individual effort.

To be sure, Hayek makes a distinction between rules of just conduct that aim at limiting 'the range of permitted action' (1973, 127) and constitutional rules of organisation that aim at allocation and limitation of powers within the organisation (1979, 134). He also recognises that 'constitutional rules have always been subject to deliberate construction' (1973, 90). This further lessens the possible contrast between the contractarian and evolutionary perspectives[11]. Sugden suggests a more moderate version of contractarianism where 'the object is to evaluate possible changes in the institutions of an existing society, using a criterion of agreement that is defined relative to the knowledge and the *conventions* that prevail in that society' (1993, 421, emphasis added). This chapter tries to demonstrate how conventions might enter into constitutional reasoning.

[10] This argument is limited to concern general rules only since it is obvious that people value freedom to experiment within appropriate or agreed rules.

[11] For a detailed account on the differences and similarities between constitutional and evolutionary perspectives see Vanberg (1983).

# 3 The constitutional theory of the firm

Firms, like other organisations (clubs, associations, states, etc.), are constituted by their members. By entering into an organisation a member becomes subject to the authority system of that organisation. An individual voluntarily gives up some of her autonomy in return for the benefit she gains from participation. When entering an organisation the individual not only accepts the authority system, but is also willing to submit part of her resources to be pooled and subjected to unitary control. It is through the exercise of control over the pooled resources that an organisation can meaningfully be treated as an acting unit. The constitution of a business firm states the terms of membership as well as the member's rights of participation in controlling the combined resources. (Vanberg and Buchanan 1986, 216) As Vanberg has pointed out, individual decisions and actions are interrelated and coordinated within an organisation to the extent that allows us to refer to concerted action (Vanberg 1994, 135). Many desirable aspects in the firm dynamics depend on the success of coordinating efforts among the members and on the ways that rights are defined and justified. Capability accumulation, knowledge creation and dissemination, communication and coordination of plans and actions are examples of such aspects. It seems reasonable that we should direct our analytical interests toward the constitutional dynamics of firms when long-term developmental issues are studied.

## *3.1 Constitution as a remedy for rent-seeking*

The constitutional rules of an organisation can be described as solving two types of problems: those arising (1) in team use of pooled resources and those arising (2) when the social product of collective endeavour is distributed among the members (Vanberg 1994, 139). The former type of problems refer to *knowledge* problems of how to arrange and coordinate various tasks within the organisation. The latter type of problems seem to correspond better with the *conflictual* aspects of self-interested members. The central criterion for agreement is that a rule needs to be general enough to facilitate impartial judgement. Rules of distribution do not provide uncertainty to the extent that the members could not foresee how their positions would be affected. With regard to privately owned business firms, an equal-share rule is not more prominent than any other alternative. One solution to alleviate conflicts of interests in distribution is to examine how far property rights can be developed to provide prominent demarcation in collective endeavour.

It appears intuitively obvious that if the property rights within a firm are ill-defined or ill-protected, the members suffer through reduced incentives to put an effort and increased incentives to rent-seeking. Gifford

is in line with other asset specificity theorists when he recognises that the core problem arises when it would be in the interest of the firm to have the members making *firm-specific investments* (1991, 91). If property rights are ill-protected, members remain vulnerable to rent-seeking on the part of others and thus remain reluctant to make such investments. The constitution of a firm is then viewed as a remedy for this undesirable state of affairs. The constitution is seen as a set of *interdependent long-term contracts* among the members (ibid., 92).

The role of the constitution, for Gifford, is to 'set up a system of constraints, limiting the ability of individuals and coalitions to impose external costs on others' (1991, 92). A constitution is thus designed primarily for limiting opportunism within organisations. For Gifford, the remedy for rent-seeking tendencies is a constitution created by the owner (or her agent) 'to maximize the sum of the present values of all the assets used in the firm' (1991, 93). The central purpose for the owner to set up constitutional constraints is to provide incentives for the employees to make firm-specific investments. This is accomplished by protecting the property rights of the employees to their firm-specific investments. 'By creating an efficient constitution the owner of the firm maximises the value of his own assets in the firm and at the same time those of the other firm members' (ibid.). The positive externalities that the owner thus creates are internalised by other firm members. This can partly explain the motivation for an individual to join a firm. A member can gain access to the pool of knowledge and is at the same time protected by the constitutional rules against potential rent-seeking by other members.

Furubotn (1988) is in the same line of reasoning as Gifford. With *codetermination*, Furubotn means a provision of control rights that give the employees who make firm-specific investments part of the firms' control rights. The decisive criterion is whether or not representatives of labour take part in the firm's decision-making processes at board level (ibid., 166). The core idea of the article is to show that the firm maximises its profits by giving those employees who make firm-specific investments their share of decision-making rights. Furubotn argues that the firm is actually a 'joint investment' among capital and labour providers and therefore the employee-investors should be regarded as equity holders (ibid., 168). The sharing of control rights via codetermination is then argued to provide some assurance that '*all* interests will be considered in decision making and that unfair allocation of quasi rents will be prevented' (ibid., 168-9, emphasis in original).

A core economic problem with relation-specific investments is the question of who should bear their costs. If the firm carries the full costs of such an investment, then the employee is not adversely affected as she receives her normal wages. But if the employee pays a firm-specific investment, then she should be compensated for it, otherwise she is not

willing to make such investments. Both Gifford and Furubotn are in this line of reasoning. Unfortunately, they do not discuss examples in detail.

Let us consider a simple but hopefully illustrative example. A firm buys a generic spreadsheet programme. A manager then wants some of her subordinates to learn to use it to improve their productivity and skills. This spreadsheet programme is so popular that most business firms use it. One can thus conclude that learning to use it is a highly non-firm-specific investment on the part of the employee. On the other hand, the programme-related tasks are both firm-specific and generic. They are generic in the sense that the employee learns about its special characteristics while she operates the programme. The firm-specific part would be the information that is managed by the programme.

How is one to interpret this imaginary example in light of the above discussion? Would it not be so that the employees who are to increase their generic desirability and skills should pay their generic share of the purchase of the programme and the training? How can we discover which part of an employee's knowledge is firm-specific and what is not? We may have a knowledge problem about the demarcation.

A potential problem with Furubotn's idea is that increasing the number of decision-makers within the firm may have some counterproductive effects as well. We may then have to assess the trade-off between increased decision-making costs and benefits received from the firm-specific investments that otherwise would not have resulted. This, of course, disregarding the fact that we may have difficulties in distinguishing between firm-specific and non-firm-specific assets. Furubotn's idea approaches the central criterion of goodness in explicit social contract as he claims that it would be beneficial for the firm if all interests of the worker-investors were taken into account. Unlike in the constitutional perspective, in Furubotn's analysis, the unanimity criterion is not a value in itself, though. The analysis is based on the claim of efficiency. What is missing from the analysis is a central aspect of agreement, namely, that taking everyone's interests into account may prove prohibitively costly when decision-making costs are taken into account (Buchanan and Tullock 1962).

Gifford's analysis recognises the central rationale for a constitution as constraining the self-interested behaviour of the firm members. However, his analysis remains somewhat distant to the subjectivist foundation of constitutional economics. The idea that a central agent would design an efficient constitution for the members of the firm to follow seems to disregard core issues in the formation of a constitution. The unanimity criterion in constitutional economics is established precisely because of the problem that we cannot know whether a collective choice is efficient or not in any other way than to what extent it corresponds with the interests of the parties. The efficiency criterion of the collective choice thus imitates that of market exchange. The observation of an actual exchange is a reliable means to see if the parties (at the moment of exchange) benefit from the trade.

# 4 Extending the constitutional theory of the firm

A constitution of a nation applies to every member of that nation, even the legal-political elite (this is at least the Western ideal of it). Things are not the same within business organisations, however. Power relations emerge not only through political processes within firms but are also part of the legal statuses of the members. An owner-manager has, in part, a different set of legal rights and obligations than an employee. Therefore, constitutional considerations within economic organisations differ from those at nation level.

Coleman has introduced useful terms that recognise these important distinctions. A constitution is *conjoint* when the beneficiaries and the targets are the same persons (1990, 327). A constitution of a Western nation is a good example of this. Although not every member of the nation participates to the same degree in the process of establishing the constitution, those who are targets, i.e., those who are constrained by the constitutional rules, and those who benefit from having a constitution are the same persons. Every member faces both costs and benefits from constitutional constraints. Cost incurs as the individual has to constrain her own action within the limits of the shared rules. Beneficial impact comes from others' similarly constrained behaviour.

A constitution is *disjoint* when the beneficiaries and the targets are not the same persons. As an extreme, those who benefit from certain rules may be completely different individuals than those who are subjected to those rules. An owner-manager of a firm may design a set of rules that constrains the actions of her subordinates but which do not concern herself. It is sensible to argue that economic organisations represent constitutions that have more disjoint characteristics than what can be found at the national level. As a first approximation, this could imply that business firms are characterised by more arbitrary rules than nations, and that subordinates within firms are subject to more coercive rules than their superiors.

**Exit as a prominent constraint**

Markets in which business firms are embedded provide prominent resolution mechanisms to the potential coercion of firms' constitutional arrangements. It is reasonable to argue that employees have better opportunities to vote by their feet, that is, to withdraw from a firm that enforces unjust rules, compared to emigrating from a society (Hirshman 1974, Wolff 1997). This fact alleviates the potential coercion within disjoint constitutions. Also, an employee's ongoing participation in a firm is taken as an *implicit consent* to the firm's constitution. Although exit is more operational when examining economic organisations, than when discussing entire societies, it is not entirely unspeculative regarding economic organisations.

Exit indicates that the participant is not satisfied with the present constitutional order, or, that a better alternative has been found. This leads to a logical inconsistency in the constitutional approach. Continued participation is assumed to reveal preferences about the organisational constitution. On the other hand, the normative individualistic foundation does not permit efficiency considerations other than those based on observed exchange. This means that when we observe two consecutive exchanges by the same actor, we cannot assess their comparative goodness based on procedural justification. They are both efficient precisely because the procedural criterion of goodness, that is, the exchange was observed. What this implies in the context where a member of a firm decides to change the employer is that both having stayed in the present company and entering into the new one enjoy equal procedural efficiency. Any attempt to argue that the exit is due to unsatisfactory constitutional order is inconsistent with the procedural criterion of goodness of constitutional economics. As was discussed in chapter 5, constitutional economics accepts the idea that each agreement is conjectural in the sense that it may become changed as circumstances call for it. Although one can argue that the change occurs because the old rule has become inefficient in the sense that the members do not perceive it advantageous any more, one cannot argue that one somehow knows the comparative efficiency of the new rule over the old rule when it was chosen. This issue is central to how we perceive the change in rules and thus cultural evolution. What I am arguing here is that, based on our limited reason and imperfect knowledge, insofar as two consecutive choices are based on voluntary exchange there is no secure way for us to measure their comparative efficiency. Consider two consecutive choices made by a single chooser, the first choice is about which car to buy and the second choice is over a range of shoes. We cannot claim to know which one of these choices was more efficient solely based on the observation of exchange. To be sure, the chooser does not know it either – in the consequential sense, that is. The car she just bought may turn out to be a catastrophe while the shoes she bought in the sales may serve her well for years to come.

Back to the cultural evolution. Insofar as rule change is about increasing efficiency in the sense that the old configuration has gone out of date, nothing is said yet about the comparative efficiency between the moment in time when the old rule was agreed upon and another when the new rule is established. This point is not just a matter of taste, it is an argument based on basic epistemic limitations. Consider the example on public drinking rules, discussed in chapter 5. In that example, I argued that the members of the community preferred the prohibition of public drinking. I also noted that we do not truly know such things, we just pretend we do. This is because when we compare two consecutive moments in time, the common sense explanation disregards the fact that between them knowledge has necessarily changed (cf. Lachmann 1976). The only way for them to

know that they actually preferred the prohibition was to go and see whether they liked the permission of drinking in public places. In relatively short periods of time the change in knowledge is perhaps not so important, but when it comes to aggregate phenomena such as cultural evolution, I do not think that we should go about disregarding epistemological issues.

Notice that the constitutional criterion of goodness is only concerned with the realisation of the members' interests at the moment of choice. The constitutional perspective does not pretend to have foresight into the degree of consistency between expectations and outcomes that eventually unfold. That is why entering a firm at $t_0$ point in time and entering into another at $t_1$ point in time deserve equal procedural efficiency. As soon as the agent enters the new firm at $t_1$ it may become clear that the previous firm was the better alternative. But to know this requires accumulation of knowledge that was not there before $t_1$.

### *4.1 Connecting coordination and PD rules*

My aim here is to discuss the connection between convention and social contract (see also, chapter 3). Constitutional analysis focuses on the principles and processes by which agreement, the ultimate individualistic criterion of goodness in collective endeavour, can be reached. I have argued that seeing the bargaining process as an exchange of commitment is important, but is does not necessarily reveal enough about how and why the participants come to find mutual benefit in such an exchange. Inferring mutual benefit from the observed market exchange does not take into account the institutional environment, which defines between acceptable and non-acceptable modes of exchange.

Conventions can help us to understand how the participants are able to arrive at a shared interpretation of mutual benefit and the domain of acceptable behaviour. The rise of contractual conventions (e.g., PD conventions) can provide an explanation as to why it is in the participants' consequential interests to comply with a social contract. And also, if we view the bargaining process in social contract against the established convention of multilateral reciprocity, the term 'bargaining' can be interpreted as meaning the searching of mutually beneficial terms of agreement.

Taking an agreement as the starting point is also logically problematic. What makes a social contract in an initial state problematic is that if the initial state does not already include some mutual expectations of reciprocity, a social contract remains unattainable. This refers to Hobbes' (1996) model of the initial state. In that model, the participants cannot resolve the first-mover paradox. In order for a protective agency to arise, the model must be extended to comprise at least bilateral reciprocity in a way presented by Nozick (1974). The reason why I use the term bilateral reciprocity rather than exchange is that although there is exchange, it is not symmetric in the

sense that all exchanging parties receive gains at the moment of exchange. They receive expectations of gains that require the keeping of promises.

For a general agreement to be binding, the members must have general expectations of multilateral reciprocity in the same way that is presented in Lewis' definition of convention. This is to say that resolution of PD problems is central to social contract. In both social contract and convention, mutual expectations of commitment become central. I am suggesting a close affinity between social contract and convention. Social contract presupposes agreement, but in order for an agent to have incentive to enter into agreement, she must have expectations on multilateral reciprocity within the group. In other words, social contract presupposes convention. Another connection between social contract and convention is that, despite dissimilarities in their emergence, their stability depends on the same factor, namely mutual expectations of future commitments.

Coordination rules are generally considered stable because they are self-enforcing and resistant to change (cf. Ullmann-Margalit 1977). Consider the classic example of which side of the road to drive on. After the ambiguity concerning the establishing process of the rule has been solved, no one may hope to gain by unilaterally defecting. Any attempt to do so would be self-destructive. PD rules do not enjoy the same stability effect because insofar as it is taken that the dominant strategy for each is to defect while the others cooperate, the rule will not become established at all. The basic model of the rational actor seems somewhat unsatisfactory since empirical findings suggest that communities in general constitute not only of formal rules enforced by a central authority, but also of systems of moral rules that are spontaneously enforced and that both solve and carry PD problems at the same time. Consider the rule of keeping promises. We may reconstruct the rationale for its rise through introspection. Although each would prefer a state where she alone was entitled to defect while all the others cooperated to a state where everyone cooperated, the general consequences of defection are apparent to all. What the members may want to do is to break the vicious circle of defection by establishing mutual commitment to keep promises. This solves PD problems the group members inflicted upon each other before they were able to foresee the benefits of constraining their personal, immediate interests. Thus, the fact that PD rules, while solving PD situations, are vulnerable to individual exploitation just like the initial PD problems were, may provide a central reason for the participants to conform to such rules. Since each participant can expect others to be sensitive to behaviour that can risk the stability of PD rules, it is in their interest not to pursue such behavioural forms.

The choice of a model of the individual becomes central when the stability of PD rules is considered. If the individual is viewed as being primarily focused on consequential issues, defection should always occur when there are reasonable expectations to get away with it unpunished. But, insofar as people find it in their interests not to defect even though

consequential considerations would suggest such a response, some additional explanations need to be taken into account. The present study views the individual as responsive not only to her interests over the consequences she expects from her behaviour, but also to her interests in the consistency of her action regarding personal and social rules. The procedural interest regarding personal rules gives rise to routine responses in situations where the individual would be able to engage in situational judgement. The procedural interest regarding social rules provides the individual's conformity without reference to comparative assessment over the expected consequences. For instance, moral (PD) rules impose expectations on the individual's behaviour, but not only due to consequential reasoning. Rule following is thus not necessarily unresponsive toward the particularities of a choice situation because of the individual's *inability* to engage in consequential assessment, but because her interest in finding the 'right way to go' is not directed to consequential considerations at all.

The stability of coordination rules has to do with the knowledge problem of society. The problem is not that much about conflicting interests of individuals but about their dissimilar interpretations of shared expectations. Lewis and Schelling have explained that coordination problems are often solved spontaneously through prominence or focal points (Schelling 1960, 68; Lewis 1969, 36). But insofar as focal points need interpretation in particular contexts, their guidance need not be unambiguous. Different rules can be applied to a single coordination problem and a single rule can provoke dissimilar interpretations. The degree of shared knowledge about the rules is crucial to the successful resolution of a coordination problem. Insofar as focal points and prominence are viewed as having already resolved the interpretation problem, the participants' interpretations of the appropriate behavioural responses are correct.

The instability problem inherent in PD rules can be viewed as the opposite to that of coordination rules. The participants can expect mutual gains from adherence to a PD rule, but need to resolve the question about how to ensure that the cooperative pattern is maintained. It is not so much about the knowledge problem than about conflicts of interest. If we accept the model of the individual who is rational enough to perceive gains from reciprocal behaviour, the instability is already alleviated to some extent. If the individual also views reputation as a valuable asset, the alleviation is further strengthened. To what extent reputation is believed to influence behaviour depends on how we look upon it. One alternative is to argue that reputation influences behaviour only to the extent that there are prospects for future encounters. A problem we have to deal with then is that individuals also perceive the future to be genuinely uncertain. Since we cannot predict when and where we can capitalise on our reputation, it is rational to accumulate it by generally adhering to PD rules.

Both gains from reciprocity and reputation accumulation are consistent with the model of the self-interested individual who is mainly motivated by consequential assessment. Individuals may also have adopted other ways to deal with deviants. One alternative is *moralistic aggression*, suggested by Trivers (1971). Moralistic aggression appears to be feasible not only in the environment where the probability of future interaction between individuals is substantial. There seem to be instances of moralistically aggressive action even in situations where the expected future benefits do not cover the incurred costs of the retaliation. Individuals who punish defectors even in cases where self-interest would suggest not doing so, may follow a hitherto successful rule which may provide adequate protection from the exploitative tendency of the other actors. Further, by signalling the defector's unreliability, the retaliator increases the severity of the punishment and hence the cost of deceptive behaviour as other members will become reluctant to interact with the defector. People may also incur material losses in order to reinforce norms of fairness, revenge, courage, cooperation and honesty (Argyrous and Sethi 1996, 480). These issues can be approached from both the consequential and the procedural perspectives. My aim is here to claim that it is often not the consequential consideration that motivates people to behave in, say, moralistic aggressively.

# 5 Connections to theories of the firm

A common denominator for theories of the firm is that they are characterised by their concern with the existence, the boundaries and the internal organisation of the firm. Another common theme is that explanations for these matters are based on efficiency considerations. The goal of the present approach is different. It discusses some foundational principles of a constitutional order within the business firm. The constitutional approach advocated here corresponds with the principles of subjectivism, which gives limited scope to derive efficiency claims. The present chapter is thus unable to assess to what extent constitutional rules of an economic organisation are efficient in some other sense than being desirable, judged by the members themselves.

The literature on the theory of the firm is expanding and it would be futile to try to discuss all various approaches in this context, especially in a way that would give any more light to the matters than has already been given by others (for a detailed discussion on various contributions see Foss, 1999). In this section, I will discuss some ideas from different approaches that are connected with the main theme of the chapter.

The issues connected with the constitutional perspective that are of interest here concern the contractual arrangements within the firm. Interesting issues arise from coordination problems as well as incentive-conflicts among the members. Transaction-cost considerations correspond with our immediate intuition as well. An economising individual will prefer more goods to less and less bads to more. It is therefore expected that people will try to organise production in ways that minimise various types of costs that necessarily arise from action. It is another thing to what extent lists of different kinds of costs take into account all relevant costs, or whether all those costs that influence choice-behaviour can even in principle be made operational (Cf. Buchanan 1969). Be that as it may, various approaches contribute to our understanding about the dynamics of economic organisations, a subject which is continuously changing as new, hitherto unperceived organisational arrangements are being created.

## *5.1 Incomplete contracting*

Incomplete contracting theories break with the Arrow-Debreu assumption of complete contracting. It strikes one as being rather realistic to assume that individuals do not know all the future contingencies which may affect the carrying out of a contract of any complexity or time span. Despite this, both the nexus of contract approach and the formal principal-agent theory are largely based on the assumption of complete contracting (Foss 1999).

*Coordination* is one of the themes around which the incomplete contracting approach rotates, beginning already with Coase's (1937) seminal contribution. Wernerfelt (1997), for example, argues that the firm exists because of its advantage in minimising *communication costs* in intrafirm relations. Herbert Simon (1951) emphasises the distinction between the employment contract and the market contract. This perspective contradicts another contractual idea, developed by Alchian and Demsetz (1972), that intrafirm contracts cannot be distinguished from market contracts. Their analysis implies that the firm is reduced to a fictitious legal entity. The constitutional perspective is founded on the recognition that intrafirm relations are essentially different from market relations. They are different enough to make the concept of *concerted action* operational within the firm (Vanberg 1994, 135). It is precisely the cooperative team dynamics, which are not decomposable into bilateral agreements among the members, that make intrafirm relations different from the market ones (see also, Coleman 1990). Simon (1951) argues that the advantage of the employment relationship over the market contract lies in its *flexibility*. After the employee has submitted to the governance structure of the firm, her action can be adapted more fully to unforeseen future contingencies. The constitutional perspective emphasises the exchange relation between the employee and the firm. The employee is willing to limit the range of her future choices within the structure of rules (as well as authority) in return for the advantages she expects to gain from the membership.

*Asset specificity* is another theme in incomplete contracting. Unlike the coordination approach, the asset specificity perspective highlights the organisational implications of *ex post* opportunism when relation-specific investments are involved (Foss 1999, 25). Williamson (1971, 1991) and his followers extensively discuss the implications of *opportunism* combined with Simon's concept of *bounded rationality* on different types of economic organisation. This approach resonates with the constitutional perspective of Gifford (1991).

Contracts of any complexity or time span remain imperfect. This is due to our ignorance about how future events will affect what is agreed upon. Despite this anomaly, the parties can agree as new events disclose that certain implicit terms are binding which thus help in mending the initial contract. In order for the implicit terms to be effective, the parties must share their meaning. Otherwise the agreement breaks down. In order to secure agreement the parties submit to conventions that bring coherence to their interpretations of implicit terms. This is to say that an underlying reason for a successful application of implicit terms and contracts can be found in conventions.

## *5.2 Spontaneous elements*

The constitutional perspective of the present study differs to some extent from that of contractarian philosophy as defined in Brennan and Buchanan (1985). The present approach takes into account, not only explicit agreements among firm members but also conventions. The perspective is related to approaches that emphasise the plurality and complexity of the relations within organisations. For instance, Herbert Simon states that

> To many persons, an organization is something that is drawn on charts or recorded in elaborate manuals of job descriptions. … In this book, the term organization refers to the complex pattern of communication and relationships in a group of human beings. This pattern provides to each member of the group a … set of stable and comprehensible expectations as to what the other members of the group are doing and how they will react to what he says and does. (Simon 1976, introduction to the third edition, as referred in Baker, Gibbons and Murphy 1997)

A number of writers within related perspectives share the understanding that implicit contracts and spontaneous procedures are essential components of organisational dynamics (see, e.g., Barnard 1938, Simon 1976, Granovetter 1985). The present study shares Barnard's view that many of the rules and practises are organisation-specific:

> [Consider] the lines of organization, the governing policies, the rules and regulations, the patterns of behavior of a specific organization. Though much of this is recorded in writing in any organization and can be studied, much is "unwritten law" and can chiefly be learned by intimate observation and experience. (Barnard 1976)

The present perspective is also related to Baker, Gibbons and Murphy's (1997) analysis of implicit contracts. They emphasise the role of management in 'the articulation of unwritten rules and codes of conduct, the development and maintenance of a reputation for abiding by these rules, and the use of subjective assessments and informal adaptation to events in the implementation of these rules' (p. 23). The present approach deviates from theirs in that the emphasis is on the role of conventions as constitutional constraints. The creation of implicit contracts is therefore not seen as being as 'conflict-laden' a process as Baker, Gibbons and Murphy suggest. Kreps' (1990) emphasis on the role of *corporate culture* gives some insight into these matters.

## *5.3 Corporate culture*

Kreps' (1990) analysis of corporate culture discusses the realm that should reasonably be related to spontaneous forces within organisations. In Kreps' terms, corporate culture consists of 'the interrelated principles' that the organisation applies and 'the means by which the principle is communicated' to say 'how things are done, and how they are meant to be done in the organization'. Because corporate culture is 'designed through time to meet unforeseen future contingencies as they arise, it will be the product of evolution inside the organization…'. (p. 93-4) Corporate culture does not only consist of the basic principles, but plays a role 'by establishing general principles that should be applied' (p. 126). This may be taken to be related to the evolutionary idea that once a convention has been established, it becomes a reference point for future development.

The reason why the employees of a firm have reason to expect authority to be used fairly is their expectation that *reputation* is considered a valuable asset (p. 92). I would suggest that reputation alone does not ensure fairness in adapting to unforeseen future contingencies. We need something that links reputation to new situations. That link is suggested to be in the form of conventions that provide shared interpretation of fairness and also a potential to establish shared reference-points for new events as they disclose themselves.

The present approach deviates from Kreps' analysis in that it does not assume any single and rigid focal principle (p. 130). When discussing the optimal size of an organisation, Kreps assumes that a corporate culture faces problems when the span of the principle is increased (p. 129). This is because the range of contingencies that the principle must cover must also increase. The applicability of the principle (or culture or contract in Kreps' terminology) becomes ambiguous when increasingly dissimilar contingencies are introduced (p. 130). A potential reason for this interpretation may be the disregard of rules in shaping interpretations of new contingencies. The essence of any rule is that it applies to a range of dissimilar events but what is equally important is that our perception of inexperienced events is based on our capacity to perceive them through *categories* of events, not as unique events as such (Hayek 1952). This alleviates the claim that when there is a gradual expansion of contingencies (organic growth of the firm) the rule necessarily becomes increasingly ambiguous. The relative rate of change between the categories of contingencies and the rule itself then becomes the key issue. External shocks aside, there is no *a priori* reason to assume that the change in a rule could not correspond with the changes in categories.

In Kreps' analysis, corporate culture seems to obtain a rather rigid interpretation. The situation is not alleviated by the use of interchangeable terms: focal principle, implicit contract and corporate culture (p. 130). If we assume only one focal principle or implicit contract applied in an organisation, there is reason to believe that, be it however clear and

prominent, it does not provide much behavioural guidance in unforeseen future contingencies. But if we assume that there are several principles, and perhaps conventions, things change. For Kreps, this is not a solution, though, as he claims that a wider range of principles 'may increase ambiguity about how any single contingency should be handled' (ibid.). The reason for Kreps' doubt may be found in his general approach to corporate culture as being constructed by purposeful design. From the constructivist perspective the working properties of new principles are always uncertain and may only confuse the members of an organisation. My suggestion should at this point be rather obvious. I view corporate culture as being constituted by a system of conventions as well as designed principles. Conventions facilitate a wider range of principles without necessarily increasing ambiguity in interpreting unfolding contingencies. On the contrary, a central aspect of social rules is that they shape our interpretations of dissimilar events. Even in the purest form of situational analysis, where we negotiate a situation which we have no previous experience of, we try to form a solution by referring to elements that bear some resemblance to our existing categories of recurrent patterns. This dynamic is often overlooked resulting in an unwarranted picture of our choice processes as being distant to rule-following as a behavioural disposition.

In my terminology, corporate culture would be closer to the notion of the organisation's constitutional *order,* which, although it partly results from rules and principles of designed origin, should not be taken as fully designed. In this reconstruction, conventions play a role as well as explicitly agreed rules do in creating corporate culture.

Another difficulty arises in Kreps' analysis because of its static nature. Kreps states that 'efficiency can be increased by monitoring adherence to the principle (culture). Violation of the culture generates direct negative externalities insofar as it weakens the organization's overall reputation.' (p. 126) In Kreps' treatment, corporate culture is (nearly) tangible. It seems to be easy to observe when it is strengthened, as well as when it is weakened. Both violations of the culture and their consequences seem to be readily measurable. Insofar as we remain in static analysis, corporate culture remains unaltered when all the parties follow it. Kreps claims that '[r]ewarding good outcomes that involve violations of the culture generates negative externalities [because it] weakens individual incentives to follow the principle and thus increases (potentially) the costs of monitoring and control' (ibid.). The static perspective of Kreps' analysis makes changes in corporate culture unfeasible. Any experimental activity is *a priori* announced detrimental.

In the present approach, experimental activity is central to the notion of change in social affairs. Although both coordination and PD conventions are unresponsive to situational variations, they will not remain unaltered. Even technical standards, which may, for a period of time, preclude alternative arrangements from emerging, will eventually give way to

something new[12]. In order for a convention to change spontaneously, somebody may initiate change by violating the existing convention. The violation does not have to be dramatic in the sense that it may still be based on some other convention, such as general reciprocity, and receive its justification from that. Also, existing conventions may promote the emergence of new alternatives[13].

### *5.4 The subjective – objective continuum*

The extended constitutional perspective breaks with most of the theory of the firm contributions in that it is not directed to particular outcomes in the sense of considering certain institutional arrangements as objectively efficient. It is unable to define efficiency in other than subjectivist terms, extended by the objective flavour of convention[14].

The present connection between evolution and purposeful design is reminiscent of the approach to evolution as *purposeful selection*, provided by Commons (1924). A central idea in purposeful selection is that the evolutionary process can be guided without wanting to direct it toward a predetermined goal (Vanberg 1997, 113). Evolution can thus be 'cultivated' towards an overall direction desirable from the human perspective. Vihanto (1993) is in the Hayekian line of reasoning as he recognises that the search for a 'good' society is essentially about creating favourable conditions for future discoveries, rather than choosing among existing alternatives (p. 66). The purposeful selection theme is persuasive as it combines the open-endedness of cultural evolution with the purposefulness of human beings.

In this chapter, conventions enter into constitutional considerations in two ways: (1) their formation as an implicit social contract and (2) by providing a reference point for institutional alterations. The formation of a convention presupposes, to some extent, rational calculation by the individuals involved. Individuals are taken to be rational enough to consider not only the immediate gains but are also able to perceive future benefits from constrained behaviour. The formation process does not presuppose

---

[12] See, e.g., Constant, Edward W. II (1980) *The Origins of the Turbojet Revolution*. Baltimore and London: The Johns Hopkins University Press.

[13] Think about an organisation where it has become conventional to maintain the team spirit by arranging happenings outside the working hours. The informal enforcement of the convention makes it undesirable for any single member to start complaining that she is not paid for the time she sacrifices to these events. Although there may exist no formal obligation to participate, the members are reluctant not to do so as they anticipate loss in reputation as reciprocal, participatory member if they defected. It is conceivable that the convention of collective happenings may trigger other alternatives that operate in the same direction, strengthening team relations. This idea emphasises the non-conflictual change of conventions. A change does not have to result from the violation of an existing convention. Instead, a new convention may grow from the same foundation and eventually replace the old one.

[14] Objectiveness is here considered in an open-ended, Popperian sense (Popper 1972).

that the behaviour of all the members is based on rational calculation, though. Some members may start to imitate behavioural patterns as soon as they perceive changes in their reference points. This is to say that there probably are leadership effects operating in the formation process. A highly regarded person may have become a reference point that affects the behaviour of some other members. This gives good reason not to consider the rise of a social contract or a convention purified from power relations within any group.

Arguments in theories of the firm usually receive their justification in some form of an efficiency consideration. It is *theoretically* clear that if, for example, property rights within a firm are so defined that each member bears all the costs and receives all the benefits of her actions, the system resembles the market process and can be considered to be operating efficiently (disregarding the costs of the property rights framework itself). Or, in the same vein, if through vertical integration, i.e., through restructuring property rights, the parties can resolve *ex post* opportunism, the outcome is beneficial.

The point I want to address is not so much about the theoretical treatment of efficiency. It is about transferring efficiency into empirical considerations. For instance, vertical integration is an efficient solution insofar as the costs in any real world case correspond with those defined in the theory as relevant. The problem is that although the theory may provide a list of relevant costs, we still have to connect those costs to real events. Efficiency enters the neoclassical framework as an objective notion because of the assumption of a competitive market environment. In such an environment, individuals' subjective assessments become secondary as it is believed that errors in evaluation will be, through natural selection, corrected[15].

Leaving natural selection aside, conventions play a role in establishing accounting procedures that are obviously needed when the idea of objective efficiency is transferred into reality. We can discover whether an organisational (re)arrangement is beneficial or not by examining *numéraire* flows based on existing accounting conventions. An open question remains, however, about the time span that should be considered relevant in assessing efficiency. There is good reason to assume that such an assessment remains subjective insofar as conventions on it are not established. Another matter that cannot be solved by reference to convention as an objective benchmark is the subjectivity of opportunity cost. Accounting conventions cannot resolve this problem because the data that enters the decision model is not

[15] In the theory of the firm literature, natural selection links to Darwin's later aspirations to show that evolutionary development can be viewed, not only as a struggle for existence but as a survival of the fittest. Darwin's initial theory of the evolutionary principles facilitated only adaptation to changing local environments and did not contain the later modification about evolution as a *progressive* process (for a psychohistorical account on the potential rationales for Darwin's dichotomy, see, Gould, Stephen Jay (1996) *Life's Grandeur – The Spread of Excellence from Plato to Darwin*. London: Jonathan Cape.).

objective. Although a convention may provide guidance on how to deal with a particular factor, it cannot define its value. The subjectivity of opportunity cost lies in that *the foregone alternatives will never materialise* (Buchanan 1969). About them we can only have subjective expectations.

# 6 Conclusions

The broad goal of the chapter has been to promote an understanding of the linkage between rational constructivist and spontaneous elements in the business firm's constitutional dynamics. Business firms as voluntary organisations embody much of the same dynamics as larger organisations such as nation states. On the other hand, it is clear that the interrelations among the members of business firms are distinguishable in many aspects from those among the members of a society. The broadly defined constitutional perspective can provide an explanation for institutional change within firms without introducing non-individualistically definable efficiency criteria, or without assuming away the subjective elements by referring to natural selection.

A central reason for my attempt to extend the constitutional approach to the firm by conventions is a conjecture that unwritten rules, or shared interpretations of coded rules, often bear more behavioural impact than the coded rules themselves. Within economic organisations the mediation of conflicts resembles that of common law in the sense that issues are dealt with on a case-by-case basis, but from these separate cases a body of conventions emerges and justification is sought based on precedents and analogue.

Why not then treat such a body of conventions as an implicit constitution? The implicit agreement idea in social contract is beneficial in that it permits the idea of non-verbal contracting among the participants. The spontaneous accumulation of precedents in the evolution of conventions is different from a general implicit agreement in that the scope of an agreement (or resolution) may be limited to the parties that sought the resolution. Other organisation members need not be involved in any way in the process and thus reference to implicit agreement among them is unnecessary. Furthermore, even though the process of precedent accumulation does not have to affect more than a handful of participants at a time, each resolution bears an indirect consequence to the whole organisation. The aim is to resolve the case at hand, but as a consequence the resolution becomes a potential precedent through which other cases that somehow bear resemblance to it become interpreted.

Thus, the present approach to the organisational constitution emphasises the interpretation process. The body of conventions and (implicit and explicit) social contracts are interrelated, though. Through the process of case-by-case interpretations a body of rules emerge which each participant then assesses as to whether or not she is willing to conform to such rules. Thus, conventions and implicit agreement are interrelated processes.

The common-law type of convention development in economic organisations may partly explain why the 'real' constitutions in organisations

remain heterogeneous. By real I mean the observable interpretations of constitutional rules. For instance, compensating and rewarding organisation members for their efforts is a clear and simple principle, but what particular configurations and interpretations terms like effort receives obviously varies. The institutional environment of an organisation does not provide an explanation for firms that are under the same legal regime. It is difficult to see that individual preferences would bear the central burden in maintaining the heterogeneity of constitutions either. These issues require further, empirical investigation before more can be said here.

# Chapter 8

# Constitutional Dynamics of the Open Source Software Development

# 2 Introduction

This chapter aims at illustrating some central issues examined in the study. Open source software development provides an interesting subject because of its peculiar constitutional framework. The initial social contract precludes anyone from copyrighting open source software. Therefore, conventions have risen to fill the 'gaps' in property rights. The focus will be here on the interplay among social contracts, conventions, and open source organisations.

A question may be raised about whether or not open source software projects can be seen as economic organisations. In the introductory chapter economic organisations were described as corporate actors that are defined by the following features: the members combine certain resources that are used jointly subject to certain procedural rules. These procedural rules provide the common denominator that coordinates organisational interaction among the members and can be seen as a constitution (cf. Coleman 1974, 1986, 1990). The emphasis on a constitution as the coordinating device provides degrees of freedom in interpreting what is and what is not counted as an economic organisation. Also, the term 'corporate actor' has two important connotations. The 'actor' part refers to coordinated activities to the degree that the organisation can be seen as being based on concerted action of the participants (Vanberg 1992). The 'corporate' part refers to the particular legal configuration that is observed in a group. *Merriam Webster's Collegiate Dictionary* defines the corporation as 'a body formed and authorized by law to act as a single person although constituted by one or more persons and legally endowed with various rights and duties including the capacity of succession'.

If we add an open system definition of organisations, provided by Scott (1992), an interpretation emerges that can be reflected upon when analysing open source development. Scott's open system definition explains that organisations are 'systems of interdependent activities linking shifting coalitions of participants; the systems are embedded in — dependent on continuing exchanges with and constituted by — the environments in which they operate' (p. 25). What the corporate actor and the open system perspectives share is the assumption that organisational participants need not — and do not — hold common goals.

The extent to which open source projects can be seen as corporations seems to vary. There are several companies, such as Caldera Cystems, Linux Mandrake, Red Hat, Slackware, SuSe, Turbo Linux, Yellow Dog, etc., which are corporations in the legal sense. Then again, many of the projects and distributions are not registered as companies. I shall suggest here that whether or not distributions are registered as companies is not pivotal to the discussion that follows. The aim of this chapter is to illustrate some interesting and central features of open source projects that can promote

our understanding of organisational dynamics in general, rather than to argue for a reinterpretation of economic organisations.

The open source software model is based on the freedom to use, copy, modify and redistribute software. The term *open source* means that the source code needed to modify software is provided, and that the users/developers have the right not only to use, but also to modify and distribute modified versions. The starting point is that nobody is permitted to pronounce an exclusive property right to open source software. The proprietary model with which the open source model is convenient to be compared is based on a more conventional idea of copyright. The developer/distributor reserves all rights to copy, modify and distribute while users have only the right to use the software.

The sketch of the complex and interdependent model is as follows. The elements of the model are examined in terms of their degree of intentional design vs. unintended impact, as well as in relation to their degree of importance or necessity to the process. The analysis will begin by looking at general conventions of fairness among the software-developing community. These conventions are unintended from the open source software development point of view. The conventions of fairness give rise to specific conventions of 'property' in open source development. Drawing upon these conventions, the central players in open source development designed a social contract to maintain the beneficial pattern of cooperation among developers.

Open source software itself brings important elements to the model as well. Three elements are considered here: (1) technological modularity, which is viewed here as comprising both intentional and unintended elements, (2) the objective knowledge aspect of the software itself as an enhancement to communication (an unintended element), and (3) the management of the selection process of software improvements, which is an intended element in the model. All these elements together give rise to *interests* and *capability* of the members to participate in the development of open source software. Genuine uncertainty of the overall interplay between these elements is described in a statement by Linus Torvalds, the founder of Linux, the prominent open source operating system: 'Only afterwards have we started thinking about what went right in the process' (Wow 1st June 2000).

This chapter is organised as follows. I begin by describing some essential aspects of open source software development. The second section will examine open source conventions and the social contract, together with some central reasons for their emergence and enforcement. In the third section, I examine the technology-, communication- and management-related aspects of open source development. The fourth section will examine the interplay among the spontaneous and the purposefully designed elements of the model. A central issue for the development of the open

source paradigm appears to be the question about which of the two alternative social contracts that are discussed here will become stabilised.

# 3 Open source development — aims and rationales

Open source software development is looked upon today with increasing astonishment. From the consequential point of view, it should not exist or at least not spread as fast as it does. Acquiring, developing and distributing open software is free of charge. The developers do not receive the right to own their contribution and are required to provide access for anybody to obtain their contribution. Access to and distribution of software is facilitated by modern technology, especially by the Internet and e-mail news groups.

The beginning of the open source movement, in the early 1980s, was a conscious attempt to continue the software-sharing conventions of the software developers' community. Sharing and exchanging software freely among the developers was the convention before; in the early 1980s prominent university laboratories and companies started using nondisclosure agreements to prevent the distribution of free copies (Stallman 1999). The software-sharing convention at that time was rational from the developer's point of view, as income streams were not connected to choices whether or not to distribute copies and modifications. The game was reciprocal where everyone gained by helping and receiving help from others. But the game can go on only as long as copyrights and licenses do not prevent it — and they started to do precisely that.

There were many reasonable reasons for the increasing use of copyright and restrictive licenses in the 1980s. Without going too deep into that line of development, one can hypothesise that the change from huge central computers toward personal computers was an important factor in the development. The rise of proprietary software made some members of the software developers' community uncomfortable. The question was not so much about whether it was morally correct for somebody to make money out of developing and selling useful software. It was perhaps more about how they perceived software in general. They viewed software as a general means to help people – very much like language. Nobody would like to see our common language being closed up by someone who would then have the sole right to modify and distribute it.

The open source movement arose as a countermovement to the proprietary model. In order to be able to resist the increasing dissemination of proprietary software, open source developers needed to create their own operating system, and the 'GNU' project was born (Stallman 1999). The GNU project was built upon a set of principles that can be viewed as the *social contract* of open source movement. The terms of this social contract, called *Copyleft*, were later on considered too extreme by developers who saw that in order to attract the attention of business people, they need to alleviate/omit some terms to facilitate the combination of the open source

and the proprietary models. This process appears to be increasingly in the core of open source software development today.

A distinctive organisational aspect of open source software development is that there are no predefined boundaries to an open source software organisation. Membership in a project is based on self-selection where those developers who feel capable of contribution do. An open source software project uses software development capabilities throughout the world. Suggested improvements and modifications are then reviewed by a central agency, the project management, which has the right to select between beneficial and less beneficial suggestions. The Internet functions as a prominent means of coordination and communication among developers.

A central distinction between open source and proprietary approaches in software development is that the proprietary approach allows the developers to collect rent from the secret bits of their software, while on the other hand, it forecloses the possibility of truly independent peer review. The open source approach sets up conditions for independent peer review, but precludes the extraction of rent from the secret bits (Raymond 1999).

The choice (social contract) of precluding exclusive property rights provides particular incentives to contribute to the software development, but it also selects out some more conventional incentives. It appears rational for a developer, who values signalling her expertise and cumulating reputation, to contribute to an open source software project, even without monetary reward. But if we limit the repertoire of incentives to the pecuniary ones only, it may become difficult to understand why open source projects emerge and exist. It seems to be precisely the preclusion of the possibility to extract rents that makes us wonder how such a model can work at all in the real world which is assumed to be dominated by pecuniary incentives. The basic rational choice model is silent about what particular preferences people hold and what incentives they face in particular situations.

If we truly accepted this basic assumption, we would have no apparent reason to question the applicability of the open source model based on particular incentives. Reality proves otherwise, though. The open source model is criticised because it fails to comply, not to the rational choice model, but to the conventional expectations about what makes people tick. What this amounts to is that although we seem to accept the rational choice model (when being explicitly asked about it), we tend to disregard its central limitation: the model is confined to the logic that people prefer better to worse. The astonishment that surrounds the open source model reveals an important aspect about the use of rational choice theory: when we refer to the rational choice model, we actually *mean* something more than what the model provides us. We are not persuaded by the empirical evidence that open source development is real, that it exists. There is something suspicious about its existence because what we really mean by referring to the rational choice model is that only certain types of incentives should be

considered, those that are conventionally thought to motivate people. This is to say that assumptions about proper incentives play a central role in economic modelling.

A central issue that open source software developers need to tackle is the special structure of rights and responsibilities. The rejection of the conventional property rights structure complicates the accountability of each developer. As the social contract does not encourage demarcation of various rights among developers, conventions emerge to remedy the situation. A (potentially tautological) hypothesis can be put forward: where there is a clear discrepancy between the need for demarcation between different rights and the existing structure of rights, extensions and modifications to the existing structure tend to happen, springing from the existing bodies of legal traditions.

Open source development benefited from building upon conventions that had been developed in software-developers' communities earlier. The open source conventions need not be discovered in the genuine sense because for those who shared the earlier cooperative behavioural pattern, they are rather obvious remedies to the problems that would predictably arise in their absence. Thus, the existence of a stabilised PD convention of multilateral reciprocity among software developers influence their procedural interests to continue cooperating even if the property rights in the software world were changing toward the proprietary model.

# 4 Open source social contracts and conventions

Open source software development is based on a peculiar pattern of rights and responsibilities. This section will analyse the terms of the open source social contract, which was intentionally designed to preserve open development, and the conventions that arose to frame this development. The social contract prevents anyone from pronouncing exclusive property rights to open source software, whereas central open source conventions function precisely to define particular property rights. There is an interesting interplay between the deliberate aim of the social contract and conventions that define boundaries between acceptable and unacceptable behaviour.

To complicate things, there is an additional set of principles, the Open Source Definition (OSD) (appendix B), which was designed to provide more closure than the social contract, the Copyleft. A number of licenses have emerged based upon the OSD. The development of those licenses shows a tendency away from the original social contract towards a hybrid version of open source and proprietary principles.

**Copyleft and GPL — the original social contract**

The aim of the open source movement was to counterbalance the increasingly proprietary world of software development. In order to secure that open source software, after having left the hands of the original developer, remains open source, a legally binding set of rules needed to be established. The solution was found in the combination of copyrighting and licensing. Copyright resolved a problem that, e.g., public domain software suffered from. Public domain software is free in the extreme sense that anyone is free to take a copy of such software, pronounce it as her own, change the author (or any other) information, and start selling it under whatever license she wishes.

The open source people were knowledgeable of the risks that complete freedom might bring about (such as converting open source development into closed source), so they chose to copyright their software, and to provide the General Public License (GPL) (appendix A) based upon the principles of Copyleft to go with it. Copyleft uses copyright law but functions as the mirror image of the conventional use of copyright. The central idea of Copyleft is to give everyone permission to run a programme, to copy, redistribute, modify, and distribute modified versions — but not the permission to add restrictions to the license. It is important to notice that the freedom Copyleft provides does not have anything to do with price. Anyone is free to charge anything one wishes from (re)distribution — as long as the same opportunity is open to anyone else as well.

The central aim of Copyleft being the prevention of open source software from becoming converted into closed source, some important,

although unintended, implications follow. A central license design problem is that the designer must not only consider various activities a licensee is prevented from doing, but she must also imagine various ways a licensee could circumvent any of the license terms. The aim of the GPL license is not to prevent people from distributing GPLed software together with closed source software using the same medium (such as a CD-rom). To be sure, the open source principle would have nothing against combining open and closed source software into an aggregate programme, if it were possible to demarcate where one license starts and another ends. This is, however, technically next to impossible and would provide ample opportunities for the more restrictive license to encompass the less restrictive, the end result being that the whole programme would be interpreted through the more restrictive license.

For this reason, GPL contains a term that permits distribution only as 'independent and separate works' with software based on a license more restrictive than the GPL. An attempt to combine GPLed software with another based on a more restrictive license is legitimate only if the resulting whole becomes GPLed. This is why GPL is considered viral or contagious. But we need to recognise the motivation behind this viral nature. The clause is there to protect the less restrictive license from being interpreted through the more restrictive, in other words, it prevents GPLed software from being hijacked by closed source software. I will turn to this point below when the more relaxed Open Source Definition is discussed.

### Open Source Definition (OSD) — the revised social contract

Open Source Definition (OSD) (appendix B) is a bill of rights for the recipient of open source software. It functions as a certificate that ensures that licenses accepted by OSD meet the required criteria and can thus be defined as open source licenses (Perens 1999). OSD grew from a certain degree of discomfort with the demand of symmetry and reciprocity in Copyleft and GPL. The developers of OSD wanted to better be able to connect with the closed source world and still ensure that open source software remains open source. Here are the OSD terms and a short analysis on their function:

1. *Free redistribution:* a license based on OSD may not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programmes from several different sources. The license may not require a royalty or other fee for such a sale. The rationale behind this clause is to promote free redistribution by eliminating incentives for extracting rents on others' work. This clause has the effect of retaining the game cooperative.

2. *The source code must be included*. This clause enhances the development of open source software as modifications are often impossible without having access to the source code.
3. *Derived works*: a license must allow modifications and derived works to the original software, and must allow them to be distributed under the same terms as the license of the original software. For rapid development of software, people need to be able to experiment with and redistribute modifications. This clause has an interesting implication, as it does not require any producer of a derived work to use the same license terms as the original, it only provides an option to do so.
4. *Integrity of the author's source code*: a license must explicitly permit distribution of software built from modified source code and it may require derived works to carry a different name or version number from the original software. This clause enhances reputation building among developers. People need to know who is responsible for particular modifications. The term also facilitates the distinction between official and unofficial changes to software.
5. *No discrimination against persons or groups*. This clause is based on the recognition of the Hayekian problem of dispersed knowledge. Promoting diversity of people and groups equally eligible to contribute is viewed beneficial because we do not know beforehand who will discover something valuable.
6. *No discrimination against fields of endeavour*: for example, a license may not restrict software from being used in a business. This clause encourages commercial use of open source software.
7. *Distribution of license*: rights attached to a programme must apply to all to whom the programme is redistributed without the need for execution of an additional license by those parties. This clause prevents any attempt to indirectly close up software, such as requiring a non-disclosure agreement.
8. *The license must not be specific to a product*: rights attached to a programme must not depend on the programme's being part of a particular software distribution. This clause facilitates extracting open source software from any distribution, and preserving the extracted software with the same rights as those that are granted in conjunction with the original software distribution.
9. *The license must not contaminate other software*: a license must not place restrictions on other software that is distributed along with the licensed software. This clause facilitates the *distribution* of open source software along with proprietary software, but at the same time it restricts *combining* open

source software with proprietary software under the license of the latter. So, any combined work needs to be distributed under OSD.

The third clause on derivative works contradicts the terms of Copyleft and the GPL insofar as more restrictive terms can be introduced to the modification. What this clause does is that it opens up the possibility to privatise modifications and charge money from their use. The OSD conformant BSD license (appendix C) provides precisely this. However, OSD restricts charging money from the initial license only, so the holder of the initial license is restricted from charging anything from the subsequent redistributions. This creates a tendency for the price of BSD licensed software to approach zero, but it also permits converting a derivative work on BSD software into closed source.

The critical point in preserving open source development open also in the future appears to be the modifiability of license terms. The GPL license terms themselves are outside the rights that the license provides, that is, the GPL defines rights to software which does not include the license itself. By this it prevents any attempt to modify the license terms and can thus guarantee that software which is initially distributed under GPL also remains under it, irrespective of how much it will be modified during the development. The modifiability of the BSD license terms does not provide any guarantee of the future development of open source and is thus vulnerable for rent seeking and PD dynamics to enter the game.

### *4.1 Open source conventions*

Open source conventions are based on fairness, non-discrimination and equal treatment of all parties. While most open source developers do not object to others profiting from their contribution, most also demand that no party be in an exclusive position to extract profits. A developer is willing to let someone else to profit by selling her software or patches, but only as long as the developer herself could also potentially do so (consistent with both Copyleft and OSD).

Developers have observed that licenses that include restrictions on and fees for commercial use have serious chilling effects. Restrictions on use, sale, modification, or distribution inflict cost of conformance tracking and, as the number of packages people deal with rises, uncertainty and potential legal risk increases. This outcome is considered harmful, and there is therefore social pressure to keep licenses simple and free of restrictions. Despite this convention, new variants of more restrictive licenses have been developed (such as the BSD). A potential source for this development are aspirations to benefit from the available open source software together with the positive value of the open source label, and at the same time to gain a monopoly position through exclusive rights to software.

A central function of open source conventions has to do with preserving the peer-review culture based on multilateral reciprocity. License restrictions designed to protect intellectual property or capture direct sale value often have the effect of making it legally impossible to fork[16] the project. While forking is considered a last resort, it is considered critically important that that last resort be present as protection against a maintainer's incompetence or defection (Raymond 1999).

The open source social contracts (both the Copyleft and the OSD) permit that anyone can hack anything. Nothing prevents half a dozen different people from taking any given open source product, duplicating the sources, running off with them in different evolutionary directions, all claiming to be 'The' product. In practice, however, such forking almost never happens. Splits in major projects have been rare, and always accompanied by re-labelling and a large volume of public self-justification. The open source movement has an elaborate but largely spontaneous set of ownership conventions. These conventions regulate who can modify software, the circumstances under which it can be modified, and who has the right to redistribute modified versions back to the community (Raymond 1998):

- There is strong social pressure against forking projects. Forking does not happen except under special conditions, with much public self-justification, and with a renaming.
- Distributing changes to a project without the cooperation of the moderators is disapproved.
- Removing a developer's name from a project history, credits or maintainer list is not permitted without the person's explicit consent.

What does 'ownership' mean when property is infinitely reduplicable, highly malleable, and there are no explicit coercive power relationships in the surrounding culture? The owner(s) of an open source software project are those who have the exclusive right, recognised by the community at large, to redistribute modified versions. According to the standard open source licenses, all parties are equals in the evolutionary game. But in practice there is a well-recognised distinction between 'official' patches, approved and integrated into the evolving software by the publicly recognised maintainers, and 'rogue' patches by third parties. Rogue patches are unusual, and generally not trusted (Raymond 1998).

Conventions encourage people to modify software for personal use when necessary. Conventions are also rather indifferent to activities of redistributing modified versions within a closed user or development group.

[16] Forking means to take any given open source product, to duplicate the sources, and to develop them in different evolutionary directions.

It is only when modifications are posted to the open source community in general, to compete with the original, that ownership becomes an issue.

There are, in general, three ways to acquire ownership of an open source project. One is to set up a project. When a project has only had one maintainer since the beginning and the maintainer is still active, convention does not even permit a question as to who owns the project. The second way is to have ownership of a project to be transferred by the previous owner. There is a clear convention that project owners have a duty to pass projects on to competent successors when they are no longer willing or able to invest needed time in development or maintenance work. The third way to acquire ownership of a project is to observe that it needs work and the owner has disappeared or lost interest. The responsibility of the acquirer is to make an effort to find the previous owner. If the previous owner cannot be found, then the acquirer may announce in a relevant place (such as a Usenet newsgroup dedicated to the application area) that the project appears to be orphaned, and that she is considering taking responsibility for it. Convention demands that the acquirer allow some time to pass after the announcement. In this interval, if someone else announces that they have been actually working on the project, their claim exceeds the newcomers. It is considered good form to give public notice of the intentions more than once.

These features suggest that the conventions are not accidental, although they may be spontaneous responses to the social contracts that do not clearly define property rights among the developers; spontaneous in the sense that such conventions are increasingly conformed to within a group facing such a social contract. Later on in this chapter the open source conventions are examined against the background of an ancient body of natural law. That discussion will demonstrate a central argument of this thesis, the emergence of conventions require, not only prominent precedents, but also interpretation.

# 5 Technology, management and communication

In this section, three further components of the model are introduced: (1) communication, (2) technological modularity, and (3) project management in open source software development. The central aspect in the communication structure considered here is unintended, it has to do with the nature of software *per se*. The technological modularity demonstrates both intentional and unintended aspects of open source development, and the same goes for project management.

**Communication and objective knowledge**

At first glance, concepts like *informal networks* or *communities of practice* seem to illustrate well what is going on in open source software organisations. A well functioning organisation needs appropriate means for communication and knowledge sharing among its members. Whenever informal networks appear, they tend to generate their own norms and conventions to facilitate communication, thus constituting communities of practice (Crane 1972, Lave and Wenger 1991). This happens both within and across organisations.

The development of structure in a community of practice depends on the overall *size* of the community and on the *diversity* of skills available. Collaborative performance enhancement depends not only on these two factors but also on the *rates* at which the members produce results that are beneficial for the whole community (Huberman and Hogg 1995, 74). Huberman and Hogg advocate an idea of a natural limit, or bandwidth, to the number of people an individual member can interact with in a network. This limit ranges from types of situations where the members can interact with everybody very rarely to types where a limited number of members interact very often.

Open source software projects can be analysed, however, through an alternative model of communication, which is less limited by the natural bandwidth effect. It differs from the basic network model in that the members need not interact directly with each other. There is a component that facilitates the flow of knowledge beyond what the members could attain when interacting directly with each other. This component is the *objective knowledge* inherent in the software itself (Cf. Popper 1972). What makes knowledge within an open source project so unique is the *source code* that is always provided together with the binary version.

Consider two software developers who try to communicate some functionality problems in a closed source programme, say Microsoft Word. Neither of them has the access to the source code as they do not work for Microsoft. When they discuss the problem they need to continuously interpret and reinterpret what the other party is saying and meaning because

they lack an exact language that would require little or no interpretation. The source code provides precisely that function in two distinguishable ways: (1) by being an exact *language*, and (2) by being *objective knowledge* by which developers can coordinate (through trial and error) their subjective knowledge. Language can be viewed as part of the body of objective knowledge, but here language is discussed as the meaning of a *means* of communication, separate from the knowledge content of any particular sentence. This distinction can be found in, e.g., computer languages that can function simultaneously as a shared language among software developers (coordinator of meanings) and as carrying out objectively existing functions (a piece of code has an effect in software disregarding how it is interpreted).

To see the difference between the network model and the one suggested here consider the following figure 8.1:

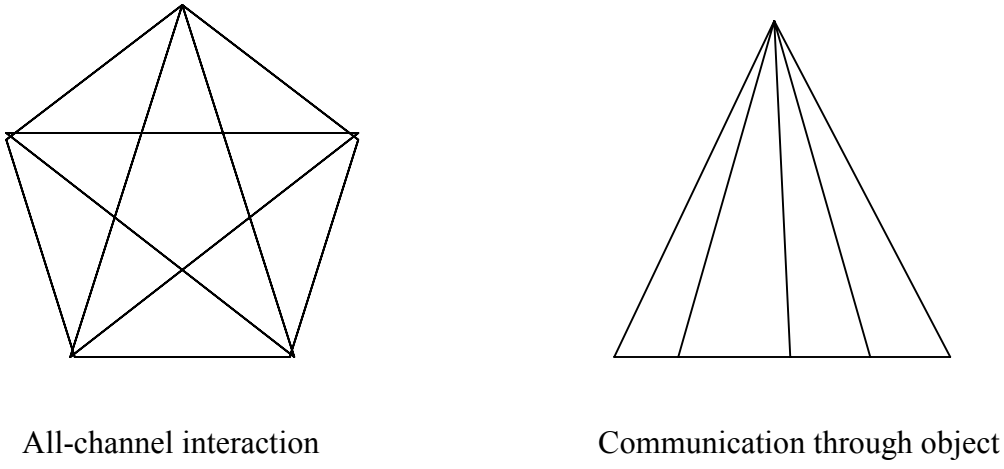


Figure 8.1: Communication models

Here we have two communication models among five members. In the first model, the members communicate directly with each other while in the second model an object (such as software) functions as an objective entity to which each member relates. A core difference between these models is that in the first alternative the members need to find out who knows what at each instance, whereas in the second model the objective entity coordinates the type of knowledge that is needed at each instance. In the first model, communication among the members is limited by their abilities (including the costs) of maintaining versatile connections (the bandwidth) whereas in the second model, only those members who at a particular instance perceive being able to add value to the development process do so. Open source software is rather an extreme case as it functions as an exact language and as objective knowledge at the same time.

Cusumano (1997, 9) suggests that small teams conducting complex tasks are more effective than large ones because it is easier to have good

communication and consistency of ideas among team members. Two issues are of interest here. First, the question of what is meant by good communication and consistency of ideas. Second, the issue of knowing in advance who will know something valuable in the future.

Good communication is assumed here to be directed at a target (such as software, which Cusumano's article deals with). Good communication may mean that things that are understandable by the majority or all members are communicated. Frictions in communication may be due to some members being smarter than the rest, or less smart (among many other reasons). Consistency of ideas is linked with good communication. What the members perceive as good communication can be the result of the consistency of their ideas. It is, however, not clear to what extent consistency of ideas works well as a primary criterion when complex systems are being developed. A novel idea may be in conflict with the established pattern of consistency, and thus become rejected before it is assessed to its full potential. A small team may work well in resolving *conflicting interests* among the group members, but the smaller the group the less versatile ideas it can produce.

This links us to the Hayekian knowledge problem of our ignorance of who may be in the best position in the future to resolve particular problems that we cannot anticipate in advance. If the group members are defined from the beginning, then only those discoveries can be made that are perceived by the members. If then consistency works as the moderator of ideas, only those discoveries are recognised that are consistent with the patterns that are already established. Discoveries become thus limited in two steps: first, by group size, and second, by the consistency requirement.

### Technological modularity in open source development

Cusumano (1997) describes how Microsoft makes large teams work like small teams. The core strategy is to break both the organisation and the products into subunits to facilitate coordination among the members and product components. The keyword is *modularisation*, both at the organisational and the product levels.

Modularity refers to a general set of principles for managing complexity. Modularity is attained by breaking up a complex system into discrete subunits, which can communicate with each other only through standardised interfaces within standardised architecture (Langlois 2000, 1). By doing so a development team can prevent the design process from becoming excessively complex at many levels at the same time. The keyword in modularisation is thus *standardisation* of the critical interfaces that subunits interact with. The degree of modularity in a system can be assessed by examining to what extent small changes in one part of a system lead to unpredictable outcomes in other parts of the system. If a system is decomposed into modules, then changes in one module do not affect

others. What modularity does is it breaks the interdependency among the subunits as each module interacts solely with the common interface.

Modularity within organisations can be divided into different types: modularity of the organisation itself, modularity of the products, and finally, modularity of property rights within the organisation. Langlois (2000) suggests, contrary to Sanchez and Mahoney (1996), that technological modularity does not necessarily presuppose organisational modularity. Indeed, there seems to be no compelling reason to assume that product modularity necessarily leads to organisational modularity.

Cusumano (1997) describes how Microsoft applies both organisational and technological modularity to coordinate and stabilise software development. Open source software, like Linux-derived operating systems, demonstrate a high degree of technological modularity but a lower degree of organisational modularity. According to Cusumano (1997) in large development projects in Microsoft, 'many team members create many components or features that are interdependent but difficult to define accurately in the early stages of the development cycle' (p. 10). And also that they need to continuously 'synchronize what people are doing as individuals and as members of teams working in parallel on different features, and periodically stabilize the evolving product features in increments as a project proceeds' (p. 11). The strategy is to continuously iterate among several designs, builds, and testing while developing a product (ibid.). All this seems to indicate that, contrary to Cusumano's view on modularisation in Microsoft, their product development is in fact non-modular. Modularity would prevent interdependency problems and activities resulting from these: continuous synchronisation and iteration as projects evolve.

The object oriented model of communication shown above illustrates open source software development. Consider the object being decomposed into modules each interacting with a standardised interface. In open source software development, the team responsible for developing a particular feature is not defined in the beginning of the project. Instead, the team itself evolves according to the capabilities of individual members to resolve particular problems that arise during the development. Communication among developers is facilitated through interfaces and is carried out in specific arenas (e.g., discussion groups on the Internet) for communicating particular issues. The organisation itself is non-modular in the sense that there is no team exclusively defined to various development projects. If a developer identifies the ability to contribute to a project at a specific instance, she can freely do so. A central benefit from not limiting the development team is that more discoveries and innovations arise during the development process. Another beneficial aspect of keeping the development team open is that we do not know in advance which developer might resolve a problem arising from the previous round of improvements. The development of open source software show a dramatically higher speed of improvements and debugging than what is achieved within the closed source

development (e.g., stability and speed of development of Linux vs. Microsoft Windows).

### Project management

Open source software organisations are flat, non-hierarchical systems. Project management can be distinguished from other members, however. The management normally consists of the property rights owners (defined by convention). The development of the open source operating system Linux has involved a myriad of extensions and improvements along the years, and yet its initial developer, Linus Torvalds, holds the position to unilaterally select among potential improvements. This suggests two unrelated issues: first, the strength of the property right conventions in open source development, and second, a conjecture about the relation between variation and selection in software development.

As time passes, the weight of other developers' contributions to any given project normally increases. As in the case of Linux, the initial developer may limit his tasks to almost solely selecting incoming suggestions. The principle of *prominence* may play a central role in sustaining the property rights convention. As years pass and thousands upon thousands of developers have contributed to the development, the only prominent person who stands out is the one who has held the right to select among trials.

This leads us to an interesting suggestion: prominence does not necessarily arise from the critical nature of the task, but perhaps from a simpler fact that the person who selects stands out because of her role as the initiator. The chain of thought goes something like this: empirical findings show that open source software demonstrates specific strengths over closed source alternatives. These have to do with the speed of improvement and bug fixing, reliability and stability, among other things. This being a general pattern it is hardly likely that open source project managers just happen to be superior in selecting good suggestions from bad ones. Rather, a potential explanation would be that selection is not the central problem, whereas creating variation is. An experienced developer can perhaps easily see what suggestions are worth looking into. And then, technological modularity enhances testing and assessing new variants. Creating variation is precisely what open source software development is superior in. The number of suggestions (variation) to any open source project of some interest exceeds what a coherent closed source development team could ever come up with.

This links us back to the nature of prominence in the open source property right convention. Insofar as selection is not the critical issue, but the creation of variation is, important contributions should have some role in the property right structure. The result would be that open source software would be owned by many, instead of by few. This would be dysfunctional from the project management point of view. Consider

suggestions for improvements being voted on in discussion groups. The dysfunctionality of voting assumes of course that software developed by voting would not be any better than another developed by the single selector model. The fact that voting is not generally used promotes the argument that selecting is not the central problem.

This section has suggested three central features to the open source software model: (1) acknowledgement of the dispersed nature of knowledge and of the problem of stimulating the growth of knowledge, (2) communication through an objective entity that functions as a communication interface among the members, and (3) technological modularity of the software.

# 6 Dynamics of the model

### Social contract and convention

As explained earlier in this chapter, open source social contracts (Copyleft and OSD) and conventions work in opposite directions. The social contracts facilitate open development by preventing exclusive property rights while conventions define property rights among the members. It is, however, important to notice that social contracts and conventions have a common origin, namely conventions. A social contract, while being a product of intentional deliberation, depends on conventions of fairness and just conduct. The connection becomes effective as soon as we introduce the possibility of social contract, not only to constrain behaviour, but also to modify interests. After reaching an agreement to reciprocally restrict behaviour to prevent PD dynamics from arising, the members may be better able to observe the benefits of long-term consequences. Their consequential interests toward reciprocal behaviour may increase as they learn during the game. The game becomes developmental as experience together with expectations facilitates steps to a higher level of cooperation.

### Conventions and interpretation

The development of conventions is linked with precedents and prominence (Schelling 1960, Lewis 1969). Interpreting the behavioural recommendations of conventions in specific situations may create problems even if the individual is procedurally motivated in finding the appropriate solution. The hierarchical structure of conventions does not necessarily help the task of interpretation. The individual may search for analogous conventions applied in situations somehow resembling the one at hand, or she may resort to a more general convention that applies through a number of dissimilar situations. For instance, a general convention of 'finders — keepers' that provides a moral argument for first possession is clear as a principle, but less so in empirical terms. Depending on a more precise convention of proper behaviour when finding money on the pavement, the finder may either consider herself the first possessor or not. Finding a car by the street more seldom triggers feelings of justified first possession.

Consider open source property rights conventions against the finders — keepers convention. It seems morally plausible to argue that an individual obtains the property right to an unowned resource by mixing her mental and physical labour with it (Rothbard 1982, 33). According to this Lockean idea, nobody is in the position to simply pronounce legal ownership to a vast area of land without indicating a differential relation to it by, e.g., fencing and cultivating it. Analogously, one who is the first to pick up driftwood on an unowned shore has the right to claim the ownership title to the findings

because no other principle offers more prominent justification (Sugden 1986, 95). The Western tradition of property rights is largely consistent with this principle.

The initiator of an open source software project is clearly the prominent candidate to claim ownership title to the project. The potential acquirer of an orphaned project needs to signal loudly her intentions, in order to make sure that the finders — keepers principle is applicable. In the same vein, forking is intuitively morally wrong because it violates the finders — keepers principle.

Open source development demonstrates something that seems to violate the finders-keepers principle, however. After a project has been developing for a period of time, it may turn out that someone outside the project management has put mental and physical labour into the project to a degree that might contest the right of the initial owner. The finders — keepers principle does not necessarily provide a clear-cut solution because, on the one hand, the initial owner has a strong entitlement, but on the other hand, new extensions and modifications can be viewed as new, hitherto unowned elements whose moral entitlement should go to the developer.

Examining open source conventions on ownership against the background of finders — keepers provokes a conjecture about an inherent tendency of open source development to dissolve. The realisticness of this conjecture depends on the relative strengths of finders — keepers and open source property right conventions. The inherent morale in finders — keepers deals with balancing effort with entitlement. The more effort one puts into an unowned resource, the more justified a property right claim is. The open source convention of retaining the property right with the project initiator may contradict our interpretation of justice when contributions and efforts flow from the group at large. If this is so, our interpretation of finders-keepers is closer to what I suggested above, that the creation of a modification or extension is perceived *per se* as justified basis for ownership.

Later developments in open source software suggest a tendency toward disintegration and toward the proprietary model. Instead of putting efforts to the development of one Linux operating system, the community has offered dozens of commercial Linux versions. Their prices have risen to almost the same level as Microsoft Windows, their major closed-source rival.

### *6.1 Objective knowledge, modularity and project management*

The objective knowledge aspect of open source software is clearly an unintended element. That source code functions as a coordinative language and as a functioning object at the same time, enhancing communication even though these functions have not been deliberately designed from the communication point of view.

Technological modularity demonstrates both potentially intentional and unintended elements. When Linus Torvalds in the early 1990s started

developing the Linux kernel, he probably did not have technological modularity as one of his prime goals. Technological modularity may often be the result of purposeful deliberation, but it may also grow more organically during development. Irrespective of the degree of intent, technological modularity enhances communication as developers do not have to control the whole system at once. They can focus their communication to a limited set of features they want to develop. Another communication-aiding aspect of technological modularity is the coordinative function of shared interfaces. They delimit ways of communication and reduce the demand for versatile exchange of ideas. When all parties share an interpretation of the central aspects of an interface, they do not have to test the extent to which other parties share this knowledge (disregarding the fact that discrepancies in their interpretations may occur).

The unilateral right of the project initiator to function as the sole selector seems intriguing as it does not necessarily convey the conventionally desirable picture of functional efficiency. If the conjecture of this chapter holds that selection is not the central issue in open source software development, since what matters most is the continuous inflow of variations and discoveries, then the connection between being the initiator of a project and receiving the property right to the whole project through convention appears potentially unjustified from the functional efficiency perspective.

**General and less general elements in open source software development**

Starting from a point where all the aforementioned elements are in place, what single element could be left aside and still preserve the assumably beneficial development of open source software? My answer may already be obvious at this point: the management structure. But this view is based on the assumption of the success of the conjecture that it does not matter so much who gets to select from among good and even better variations.

The central elements that open source development, or any other social development for that matter, could hardly do without are social contracts and conventions. This argument is, on the one hand, obvious as all our interaction is limited by a set of constitutional agreements and conventions, but on the other hand, it is not so obvious. The non-obviousness of social contracts and conventions has to do with the combination of their omnipresence and spontaneous nature. Only afterward can we observe that a particular contract was established or a particular convention emerged. I have tried to demonstrate that although conventions emerge from our shared interpretations of prominence and precedent, neither of them can usually provide us clear guidance in unforeseen circumstances. We still need to choose among various alternative points of prominence and interpret the connection between potential precedents.

The model of open source software development discussed here contains both universal and context-dependent elements. The dynamics of social contracts and conventions are independent of time and space, whereas open source software itself brings aspects that are less general. The objective knowledge and the technological modularity components enhance coordination and communication. An interesting question might be whether their coordinative force is decisive to the whole process. That is, could it be so that open participation would become too costly in the communication sense if these elements were not there. Conventions convey information, but their information content is not necessarily very rich. This is an essential aspect of conventions since recurrent misinterpretations, which a rich information content would bring about, harm the central function of conventions, the ability to coordinate.

# 7 Conclusions

This chapter has suggested that in open source software development conventions play an important role in defining property rights. Social contracts perform an equally important function in preventing PD dynamics from destroying the cooperative mode of interaction. The open source software itself brings elements of objective knowledge and technological modularisation that enhance communication and coordination. All these elements together reduce the need for managerial control regarding both coordination of knowledge and provision of incentives.

In this model, intentional elements do not seem to receive any apparent priority. It is recognised, however, that the design of the initial social contract plays a central role in facilitating open development. Without its restrictions to non-reciprocal behaviour, open source software would hardly have developed to what it is today. On the other hand, it is equally important to recognise the source of social contract. The designers did not genuinely discover the purpose of the Copyleft, instead, they codified something that was already there in the form of earlier conventions of the software developers' community. By setting up the Copyleft terms they wanted to continue what they perceived as beneficial development which was under attack by the introduction of the proprietary model.

The dependency of social contract upon convention becomes apparent in the establishing process of a social contract. The Copyleft would have been impossible to establish as a social contract unless the members perceived its terms as fair and beneficial to development. Although social contract is conceptually a product of intentional design, its content is so strongly based on spontaneous development of conventions that it becomes difficult to distinguish what parts of its content are *not* already established by surrounding conventions.

A central question concerning the future of open source development is: which one becomes the prevailing social contract, Copyleft or OSD? If Copyleft wins out, then open source development has better chances to remain genuinely open, at the cost of foregone profits from the proprietary model. If the OSD/BSD becomes the social contract, it may enhance the destruction of open source development because the BSD license does not prohibit changes in the license itself, even though the consequence might be the transformation from open to closed source.

This option to take open source private and make a profit has several consequences. (1) The existence of the option *per se* changes incentive structures as the members understand the dynamics of PDs. 2) Opportunities for defection lead to changes in expectation about how other members will behave in the future. (3) The changed expectations reinforce incentives to defect. The important aspect in this process of incentive change is that the triggering element does not have to be connected to real

events. The fact that an option exists, may be enough to bring reluctance toward contributing to development that is vulnerable to defection. Another important aspect in this development has to do with reference point consideration (see further ch. 2). If the Copyleft did not exist as the initial social contract, the members would not perceive OSD/BSD as a potential deterioration of cooperation. A change from a more to a less cooperative mode of interaction may occur. However, this does not provide sufficient evidence to conclude that the default response in PDs is defection.

# Chapter 9

# Conclusions

# Conclusions

This study has examined rule following as an alternative behavioural mode to situational judgement. Although situational judgement can be seen as a distinct type of choice behaviour, it is nevertheless based on rule following at some cognitive level. Hayek's theory of mind (1952) explains how perception is established through the categorising disposition of the mind, making no fundamental distinction between rule following and situational judgement.

The study has maintained that analysing rule following from the consequential perspective alone provides an incomplete picture of the choice behaviour that the individual is engaged in when deciding which rule to follow. Reference point considerations and *status quo* preference imply in the direction that the individual's choice behaviour is often better described in a non-consequential manner. The present study tries to contribute to this issue by hypothesising that while reference point and *status quo* preferences can be seen as predispositions, the procedural interests explanation requires conscious attention and evaluation in order for the agent to arrive at an appropriate interpretation of a rule in a given situation.

Rule-individualism explains the individual's choice behaviour from the consequential perspective. The result is either a second order rational choice among rules, or an emphasis on the cognitive limitations as the rationale for rule following. But since we can observe that individuals engage in situational judgement as well, the cognitive limitations do not seem to explain rule following alone. Even though the individual's cognitive capacity is limited, she uses that limited capacity to develop expectations of the consequences that alternative choice options provide. This study has argued that a central issue in differentiating between rule following and discretion are interests that can be directed either toward consequences or toward the appropriateness of behaviour regarding the rules that are judged as appropriate by the actor.

Conventions play a central role in the present analysis, and for several reasons. First of all, conventions provide an important part of the essential body of knowledge to which procedural interests are directed. This is to say that even in their private realms, individuals may apply a type of a decision process *as if* they were engaged in a social choice. The second central issue regarding conventions is that the individual may have procedural and/or consequential interests in general conformity to a convention, any convention. This is where the procedural interests come to play a role again. This study maintains that the individual, living her life under a complex web of rules, ranging from pure coordination rules to mixed-motives rules, is not willing nor able to adjust her behavioural response to perfectly match each type of situation with varying degrees of coordination and noncooperation. Instead, she coordinates her actions with others most of the time, often not recognising that a situation involves PD dynamics. *The same cognitive incapacity*

*that prevents case-by-case maximisation prevents maximising adjustment to different types of rules*[17]. This aspect goes easily unnoticed if it is assumed that a behavioural response is always defined by the *type* of rule. The picture changes, however, as soon as we assume an actor living her life under all kinds of rules, some of which represent more coordination aspects while others have more noncooperation features. In such a world choices about whether or not to conform, and when and how not to conform become much more difficult than when examining rules conceptually.

This connects us to the third aspect of conventions. Because of reasons described above the individual may have similar interests to general conformity irrespective of the varying degrees of coordination and noncooperation features of rules. This relates to the positive argument that cognitive limitations direct the individual's interests toward procedural issues. Thus, the agent is not only incapable of perfectly adjusting her behavioural response to different types of rules, but she is also *interested* in an appropriate behavioural response. And since her capacity is limited in the consequential realm her interest is directed to finding a proper rule to apply. This means that certain interests emerge as a consequence of human limits. Notice that this argument is not completely tautological because it is reasonable to argue that inabilities and limitations direct the agent to search other available behavioural modes. The cognitive limitations explanation to rule following is this type of an explanation.

The fourth aspect of conventions examined in this study relates to the structural argument of constitutional economics that evolution of rules takes place within the framework of constitutional rules, which is primarily of designed origin. Regarding the hierarchical structure of rules, such a position is as good as another stating the opposite: constitutional rules are dependent on spontaneously evolving rules that define fairness and justice at any given time. But when it comes to the logic of reasoning, the evolutionary position seems to gain priority. The ultimate criterion of goodness in social contract is voluntary agreement. Logically, agreement presupposes something without which the term would have no meaning, namely, mutual expectations[18]. Since mutual expectations are the core of conventions, it is convention that is logically prior to social contract.

Chapter 4 discusses a general issue of whether or not purely positive features exist in the social realm. The position of this study maintains that all aspects of social life are at least partly dependent on rules of the relevant group. Opportunity cost considerations need to take into account the possibility that the participants may *want* to create high opportunity costs. And irrespective of such collective aims, the way each member faces

---

[17] 'Maximising adjustment' refers here to ability to always coordinate even when there are no conflicting interests involved, and to ability to always calculate the net benefit of defection when conflicting interests are involved.

[18] If expectations of future performance are not present in the notion of agreement, the term has no behavioural influence.

opportunity costs varies across individuals. So, even though low opportunity cost is *conceptually* tempting, it does not need to be desirable for the participants as a group nor for any given participant as an individual member.

The discussion in chapter 4 about how the boundary between voluntariness and coercion suggests that even though it is not always necessary to consider how the boundary is defined, a conceptual examination reveals something that would otherwise easily go unnoticed. Even though human beings share something resembling natural rights, that is, rights that appear to be rather generally shared irrespective of cultural differences, group-dependent conventions do play a role in forming particular rights and obligations. This implies difficulties with regard to perspectives that disregard the normative content of group-dependent rules, such as, to natural selection as a globally maximising process. This issue may present problems to fields of analysis that deal, explicitly or implicitly, with institutions but whose aim is to provide positive explanations. What is an efficient firm in the USA may be rather different from an efficient configuration in Sri Lanka. A central strength of the constitutional perspective is precisely in its ability to define efficiency irrespective of cultural idiosyncracies. This contribution alone should provide enough reason for the theory of the firm literature to embrace constitutional analysis within its realm.

While the strength of the constitutional perspective is its appreciation of the normative content of rules, the perspective remains weak in another aspect. The strict criterion of goodness is problematic because, it is maintained here, it is not permissible to jump from the unanimity criterion into a sub-unanimous alternative and simultaneously retain the constitutional justification. Saying that sub-unanimous rules fulfil the constitutional criterion only because at some constitutional point in time the members decided upon a *principle* of post-constitutional rule making is not viewed satisfactory here. The principle *per se* does not distinguish between justified and unjustified rule making; what does is the process by which post-constitutional rules are decided. Insofar as a choice of the category is less than unanimous (which is expected to be the case since the consequences are assumed to be more readily assessable with post-constitutional rules), no guarantee of mutual benefit should be expected.

For this reason alone, the constitutional perspective appears to be more applicable in relatively small groups, such as business firms (compared to nations). Thus two central features of the constitutional perspective imply that it might be able to provide a beneficial method to the study of economic organisations and business firms. The individualistic foundation combined with the normative content of rules promote the idea that goodness and efficiency are necessarily issues that cannot be defined without reference to the desires of the people involved. Furthermore, the applicability aspect of the constitutional perspective implies that a closer

correspondence between individual preferences and collective decision processes can be attained in smaller groups, thus favouring such organisations as business firms.

Chapter 6 maintained that even though organisational decision-making is analysed in the light of decision rules, these contributions are silent about rules that define the basic, constitutional rights of the participants to pursue such decision-making in the first place. Constitutional rules define participation in the organisation, the right to decision-making processes, and the allocation of the organisation's outcome. These rights constitute the basic structure and the working properties of an economic organisation. Without their presence we could not perceive something to be an economic organisation.

Since the constitutional rules of an economic organisation influence what kinds of decision-making rules will be established, analysing decision-making rules alone does not provide a satisfactory view of the impact that organisational rules have upon organisational dynamics. For instance, careless (re)design of options schemes (and other rewarding schemes) in organisations may result in negative unintended consequences as the designers fail to acknowledge that such schemes carry important constitutional impact as well.

In chapter 8 the examination of open source software development provides some empirical illustrations to the issues discussed throughout the study. The constitutional rules play a central role in stabilising expectations among the participants. Interestingly enough, the explicit social contract among the participants was purposefully designed to prevent property rights from entering and interfering the mutually beneficial game. Thus, conventions emerged spontaneously to define how collective resources were to be used and how outputs were to be divided. It appears tempting to conclude that the degree of rule-governed behaviour remains more or less stable in a given culture. If formal, explicit organisational rules do not exist to bear their influence, the participants will come up with alternatives to stabilise expectations, either by agreement or by spontaneously restricting their own behaviour. Insofar as this conclusion is reasonable, the underlying explanation for such stability may be found in the preexisting conventions that provide cues to the participants as to what kind of social contracts and conventions are expectable.

# References


Alchian, A. A. and H. Demsetz (1972) Production, Information Costs and Economic Organisation. *American Economic Review*, 62, 777—95.

Argyrous, George and Rajiv Sethi (1996) The Theory of Evolution and the Evolution of Theory: Veblen's Methodology in Contemporary Perspective. *Cambridge Journal of Economics,* 1996, 20, 475—95.

Axelrod, Robert (1984) *The Evolution of Co-operation*. New York: Basic Books.

Baker, G., R. Gibbons and K. J. Murphy (1997) Implicit Contracts and the Theory of the Firm. *NBER working paper*, nr.6177. Cambridge, MA.

Barnard, Chester (1938) *The Functions of the Executive*. Cambridge, MA: Harvard University Press.

Barnard, Chester (1976) Foreword to H. A. Simon (1976).

Barry, Norman P. (1984) Unanimity, Agreement, and Liberalism. A Critique of James Buchanan's Social Philosophy. *Political Theory*, 12 (4), 579—96.

Becker, Gary (1976) *The Economic Approach to Human Behaviour*. Chicago and London: University of Chicago Press.

Block, Walter and Thomas J. DiLorenzo (2000) Is Voluntary Government Possible? A Critique of Constitutional Economics. *Journal of Institutional and Theoretical Economics*, 156, 567—82.

Brennan, G. and James M. Buchanan (1985) *The Reason of Rules*. Cambridge: Cambridge University Press.

Buchanan, James M. (1969) *Cost and Choice. An Inquiry in Economic Theory*. Chicago: Markham Publishing Company.

Buchanan, James M. (1975) *The Limits of Liberty*. Chicago: University of Chicago Press.

Buchanan, James M. (1977) *Freedom in Constitutional Contract. Perspectives of a Political Economist*. College Station: Texas A & M University Press.

Buchanan, James M. (1979) *What Should Economists Do?* Indianapolis: Liberty Fund.

Buchanan, James M. (1991) *The Economics and the Ethics of Constitutional Order*. Ann Arbor: The University of Michigan Press.

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

Burke, E. (1968) A Vindication of Natural Society, in Burke, E. (ed.) *Selected Writings and Speeches*. Gloucester MA: Peter Smith.

Camerer, Colin (1997) Progress in Behavioral Game Theory. *Journal of Economic Perspectives*, 11 (4), 167—88.

Camerer, Colin and Marc Knez (1997) Coordination in Organizations: A Game-theoretic Perspective, in Shapira, Zur (ed.) *Organizational Decision Making*, 158—88. New York: Cambridge University Press.

Coase, Ronald H. (1937) The Nature of the Firm. *Economica* 4 (November), 386—405.

Coleman, James S. (1974) *Power and the Structure of Society*. New York: W.W. Norton & Co.

Coleman, James S. (1986) *Individual Interests and Collective Action: Selected Essays*. Cambridge: Cambridge University Press.

Coleman, James S. (1990) *Foundations of Social Theory*. Cambridge: Belknap.

Commons, John R. (1924) *Legal Foundations of Capitalism*. Madison: The University of Wisconsin Press.

Crane, D. (1972) *Invisible Colleges, Diffusion of Knowledge in Scientific Communities*. Chicago: University of Chicago Press.

Cusumano, Michael A. (1997) How Microsoft Makes Large Teams Work Like Small Teams. *Sloan Management Review*, Fall, 9—20.

Cyert, Richard M. and James G. March (1963) *A Behavioural Theory of the Firm*. Cambridge MA: Blackwell Publishers.

Dosi, G., L. Marengo, A. Bassanini, M. Valente (1999) Norms as Emergent Properties of Adaptive Learning: The Case of Economic Routines. *Journal of Evolutionary Economics*, 9, 5—26.

Elster, Jon (1979) *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Elster, Jon (1986) (ed.) *Rational Choice*. Oxford: Basil Blackwell.

Elster, Jon (2000) *Ulysses Unbound*. Cambridge: Cambridge University Press.

Etzioni, Amitai (1964) *Modern Organizations*. Englewood Cliffs, N. J.: Prentice-Hall.

Etzioni, Amitai (1988) *The Moral Dimension. Toward a New Economics*. London and New York: Macmillan.

Foss, N. J. (1999) (ed.) *The Theory of the Firm: Critical Perspectives*. London: Routledge.

Frank, Robert H. (1987) Shrewdly Irrational. *Sociological Forum*, 2 (1), 21–41.

Furubotn, E. G. (1988) Codetermination and the Modern Theory of the Firm: A Property-Rights Analysis. *Journal of Business*, 61 (2), 165—81.

Gauthier, David (1967) Morality and Advantage. *Philosophical Review*, 76, 460 –75.

Gauthier, David (1998) David Hume, Contractarian, in Boucher, David and Paul Kelly (eds) *Social Justice*, 17—44. London: Routledge.

Gifford, Adam Jr. (1991) A Constitutional Interpretation of the Firm. *Public Choice*, 68, 91—106.

Gouldner, A. W. (1957) *Patterns of Industrial Bureaucracy*. Illinois: Glencoe.

Granovetter, Mark (1985) Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, 91, 481—510.

Gray, John (1984) *Hayek on Liberty*. London: Routledge.

Green, D. P. and Ian Shapiro (1994) *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven, CT: Yale University Press.

Habermas, Jürgen (1990) *Moral Consciousness and Communicative Action*. Cambridge: Polity Press.

Harper, David A. (1996) *Entrepreneurship and the Market Process. An Inquiry into the Growth of Knowledge*. London and New York: Routledge.

Hayek, F. A. (1937) Economics and Knowledge, in Hayek (1948).

Hayek, F. A. (1945) The Use of Knowledge in Society. *American Economic Review*, 35, 519—30.

Hayek, F. A. (1948) *Individualism and Economic Order*. London: Routledge & Kegan Paul.

Hayek, F. A. (1952) *The Sensory Order. An Inquiry into the Foundations of Theoretical Psychology*. London: Routledge & Kegan.

Hayek, F. A. (1960) *The Constitution of Liberty*. London: Routledge & Kegan.

Hayek, F. A. (1967) *Studies in Philosophy, Politics and Economics*. London: Routledge & Kegan.

Hayek, F. A. (1973) *Law, Legislation and Liberty, vol. 1, Rules and Order*. Chicago: The University of Chicago Press.

Hayek, F. A. (1978) *New studies in Philosophy, Politics, Economics and the History of Ideas*. London: Routledge & Kegan.

Hayek, F. A. (1979) *Law, Legislation and Liberty. Vol. 3, The Political Order of a Free People*. Chicago: The University of Chicago Press.

Hayek, F. A. (1988) *The Fatal Conceit. The Errors of Socialism*. London: Routledge.

Heiner, Ronald (1983) The Origin of Predictable Behavior. *The American Economic Review,* 73 (4), 560—95.

Helson, Harry (1964) *Adaptation Level Theory: An Experimental and Systematic Approach to Behavior*. New York: Harper & Row.

Hirschman, Albert O. (1970) *Exit, Voice, and Loyalty. Responses to Decline in Firms, Organisations, and States*. Cambridge MA: Harvard University Press.

Hobbes, Thomas (1996 [1654]) *Leviathan*. Oxford: Oxford University Press.

Hodgson, Geoffrey M. (1991) Hayek's Theory of Cultural Evolution. An Evaluation in the Light of Vanberg's Critique. *Economics and Philosophy*, 7, 67—82.

Huberman, B. A. and Tad Hogg (1995) Communities of Practice: Performance and Evolution. *Computational and Mathematical Organization Theory*, 1, 73—92.

Hume, David (1969 [1740]) *A Treatise of Human Nature*. London: Penguin Books.

Hume, David (1987) *Essays, Moral, Political, and Literary*. Indianapolis: Liberty Fund.

Kahneman, D. and Amos Tversky (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47 (2), 263–91.

Kahneman, D.; J. L. Knetsch; R. H. Thaler (1990) Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, 98, 1325—48.

Kirzner, Israel, M. (1973) *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, Israel M. (1985) *Discovery and the Capitalist Process*. The University of Chicago Press: Chicago.

Kirzner, Israel M. (1992) *The Meaning of Market Process: Essays in the Development of Modern Austrian Economics*. London: Routledge.

Kley, Roland (1994) *Hayek's Social and Political Thought*. Oxford: Clarendon Press.

Knetsch, Jack L. (1989) The Endowment Effect and Evidence of Nonreversible Indifference Curves. *American Economic Review*, 79, December, 1277—84.

Koffka, K. (1935) *Principles of Gestalt Psychology*. London: Routledge & Kegan Paul.

Kreps, D. M. (1990) Corporate Culture and Economic Theory. In: J. Alt and K. Shepsle (eds) *Perspectives on Positive Political Economy*. Cambridge University Press.

Lachmann, Ludwig M. (1976) On the Central Concept of Austrian Economics: Market Process, in Dolan, Edwin G. (ed.) *The Foundations of Modern Austrian Economics*, 126—32. Kansas City: Sheed & Ward, Inc.

Langlois, R. N. (1986) Coherence and Flexibility: Social Institutions in a World of Radical Uncertainty, in Kirzner, Israel (ed.) *Subjectivism, Intelligibility and Economic Understanding*, 171—91. London: McMillan.

Langlois, R. N. (1992) Orders and Organizations: Toward an Austrian Theory of Social Institutions. In: *Austrian Economics: Tensions and New Directions*. B. J. Caldwell and S. Boehm (eds), 165—83. Norwell, MA: Kluwer Academic Publishers.

Langlois, R. N. (1995) Do Firms Plan? *Constitutional Political Economy*, 6 (3), 247—61.

Langlois, R. N. (2000) Modularity in Technology and Organization. Paper presented at 'Austrian Economics and the Theory of the Firm' conference, August 19—17, 1999, Copenhagen Business School, second draft.

Lave, J. and E. Wenger (1991) *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Le Menestrel, Marc (1997) Consequential Rationality, Procedural Rationality, and Optimal Nash Equilibrium. INSEAD Working Paper Series 97/114/TM.

Ledyard, J. (1995) Public Goods, in Kagel, J and A. Roth (eds) *Handbook of Experimental Economics*, 111—94. Princeton, NJ: Princeton University Press.

Leibenstein, Harvey (1987) *Inside the Firm. The Inefficiencies of Hierarchy*. Cambridge, MA: Harvard University Press.

Lewis, D. (1969) *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Locke, John (1986 [1690]) *The Second Treatise on Civil Government* (first published in: Two Treatises of Government [1690]). Buffalo, NY: Prometheus Books.

March, James G. and Herbert A. Simon (1958) *Organizations*. New York: John Wiley and Sons, Inc.

March, James G. (1994) *A Primer on Decision Making*. New York: Macmillan Press.

March, James G. (1997) Understanding How Decisions Happen in Organisations, in Shapira, Zur (ed.) *Organizational Decision Making*, 158—88. New York: Cambridge University Press.

Mayhew, Anne (1987) Culture: Core Concept Under Attack. *Journal of Economic Issues*, 21, 587—603.

Maynard Smith, John (1976) Group Selection. *Quarterly Review of Biology*, 51, 277—83.

Merton, R. K. (1940) Bureaucratic Structure and Personality. *Social Forces*, 18, 560—68.

Milgram, S. (1974) *Obedience to Authority: An Experimental View*. London: Tavistock.

Milgrom, P. and J. Roberts (1992) *Economics, Organization and Management*. London: Prentice-Hall.

Miller, Gary J. (1992) *Managerial Dilemmas. The Political Economy of Hierarchy*. Cambridge: Cambridge University Press.

Mises, Ludwig von (1966 [1949]) *Human Action. A Treatise on Economics*, 3rd ed. Chicago: Contemporary Books.

Nelson, Richard R. and Sidney G. Winter (1982) *An Evolutionary Theory of Economic Change*. Cambridge MA: Harvard University Press.

Nozick, Robert (1974) *Anarchy, State, and Utopia*. New York: Basic Books.

Nozick, Robert (1993) *The Nature of Rationality*. Princeton NJ: Princeton University Press.

O'Driscoll, Gerald P. Jr. and Mario J. Rizzo (1985) *The Economics of Time & Ignorance*. Basil Blackwell: Oxford.

Parsons, Talcott (1937) *The Structure of Social Action*. New York: McCraw-Hill.

Perens, Bruce (1999) The Open Source Definition, in Di Bona, Chris and Sam Ockman (eds) *Open Sources: Voices from the Open Source Revolutio*n. O'Reilly & Associates.

Perrow, Charles (1970) *Organizational Analysis: A Sociological View*. Belmont, CA: Wadsworth.

Pfeffer, Jeffrey and Gerald R. Salancik (1974) Organizational Decision Making as a Political Process: The Case of a University Budget. *Administrative Science Quaterly,* 19, 135—51.

Pfeffer, Jeffrey and Gerald R. Salancik (1978) *The External Control of Organizations*. New York: Harper and Row.

Polanyi, Michael. (1951) *The Logic of Liberty*. Chicago: University of Chicago Press.

Polanyi, Michael (1958) *Personal Knowledge*. Chicago: University of Chicago Press.

Popper, Karl R. (1995 [1945]) *The Open Society and its Enemies*. Vol. 2. London: Routledge.

Popper, Karl R. (1979 [1972]) *Objective Knowledge. An Evolutionary Approach*. Oxford: Clarendon Press.

Rabin, M (1993) Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83, 1281—302.

Rabin, M. (1998) Psychology and Economics. *Journal of Economic Literature*, XXXVI, March, 11—46.

Rawls, John (1971) *A Theory of Justice*. Cambridge MA: Harvard University Press.

Raymond, Eric S. (1998) 'Homesteading the Noosphere', at www.tuxedo.org/~esr/writings/homesteading/homesteading.txt, retrieved 10 March 2000.

Raymond, Eric S. (1999) 'The Magic Cauldron', at www.tuxedo.org/~esr/writings/magic-cauldron/magic-cauldron.txt, retrieved 10 March 2000.

Rothbard, Murray N. (1982) *The Ethics of Liberty*. Humanities Press: Atlantic Highlands.

Rowe, Nicholas (1989) *Rules and Institutions*. Ann Arbor: University of Michigan Press.

Sally, D. (1995) *On Sympathy*. Working paper, Cornell University Johnson Graduate School of Management.

Sanchez, Ron and Joseph T. Mahoney (1996) Modularity, Flexibility, and Knowledge Management in Product and Organizational Design. *Strategic Management Journal*, 17, 63—76 (Winter Special Issue).

Schelling, Thomas (1960) *The Strategy of Conflict*. Cambridge, MA: Harvard University.

Schlicht, Ekkehart (1998) *On Custom in the Economy*. Oxford: Clarendon Press.

Schlicht, Ekkehart (1999) Aestheticism in the Theory of Custom. A paper presented at the conference 'Progress in the Study of Economic Evolution: A Systemic Perspective on Individuals, Firms and Local Systems', University of Ancona, Italy, May 20—22, 1999.

Schotter, Andrew (1981) *The Economic Theory of Social Institutions*. Cambridge: Cambridge University Press.

Scott, W. Richard (1992) *Organizations. Rational, Natural, and Open Systems*. Englewood Cliffs, N. J.: Prentice-Hall.

Selznick, Philip (1948) Foundations of the Theory of Organization. *American Sociological Review*, 13, 25—35.

Shackle, G. L. S. (1972) *Epistemics & Economics: A Critique of Economic Doctrines*. Cambridge: Cambridge University Press.

Simon, Herbert A. (1947) *Administrative Behavior*. New York: Macmillan Publishing Co.

Simon, Herbert A. (1951) A Formal Theory of the Employment Relationship. *Econometrica*, 19, 293—305.

Simon, Herbert A. (1976) From Substantive to Procedural Rationality, in Latsis, Spiro J. (ed.) *Method and Appraisal in Economics*, 129—48. Cambridge: Cambridge University Press.

Simon, Herbert A. (1978) Rationality as Process and as Product of Thought. *The American Economic Review*, 68 (2), 1—16.

Simon, Herbert A. (1979) Rational Decision Making in Business Organizations. *The American Economic Review*, 69 (4).

Skyrms, Brian (1996) *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Stallman, Richard (1999) The GNU Operating System and the Free Software Movement, in Di Bona, Chris and Sam Ockman (eds) *Open Sources: Voices from the Open Source Revolutio*n. O'Reilly & Associates.

Stigler, G. J. and Gary S. Becker (1977) De Gustibus Non Est Disputandum. *American Economic Review*, 67, 76–90.

Sugden, Robert (1986) *The Economics of Rights, Co-operation and Welfare*. Oxford: Basil Blackwell.

Sugden, Robert (1989) Spontaneous Order. *Journal of Economic Perspectives,* 3 (4), 85–97.

Sugden, Robert (1993) Normative Judgments and Spontaneous Order: The Contractarian Element in Hayek's Thought. *Constitutional Political Economy,* 4 (3), 393—424.

Trivers, Robert L. (1971) The Evolution of Reciprocal Altruism. *Quaterly Review of Biology*, 46, 35–57.

Trivers, Robert L. (1985) *Social Evolution*. Menlo Park, CA: Benjamin-Cummings.

Tullock, G. (1985) Adam Smith and the Prisoners' Dilemma. *Quaterly Journal of Economics*, 100.

Turgot, A.-R.-J. (1973) On Universal History, in Meek R. L. (ed.) *Turgot on Progress, Sociology, and Economics*, 41—118. Cambridge: Cambridge University Press.

Ullmann-Margalit, Edna (1977) *The Emergence of Norms*. Oxford: Clarendon Press.

Ullmann-Margalit, Edna (1978) Invisible-hand Explanations. *Synthese*, 39, 263—91.

Vanberg, Viktor (1983) Libertarian Evolutionism and Contractarian Constitutionalism. In: Pejovich, S (ed.) *Philosophical and Economic Foundations of Capitalism*. Toronto: Lexington Books.

Vanberg, Viktor (1985) Liberty, Efficiency and Agreement: The Normative Element in Libertarian and Contractarian Social Philosophy. Center for the Study of Market Processes Working paper Series, George Mason University, USA, 1985—18.

Vanberg, Viktor (1986a) Individual Choice and Institutional Constraints. The Normative Element in Classical and Contractarian Liberalism. *Analyse & Kritik*, 8, 113—49.

Vanberg, Viktor (1986b) Spontaneous Market Order and Social Rules: A Critique of F. A. Hayek's Theory of Cultural Evolution. *Economics and Philosophy*, 2, 75—100.

Vanberg, Viktor (1992) Organizations as Constitutional Systems. *Constitutional Political Economy*, 3, 223—53.

Vanberg, Viktor (1993) Rational Choice, Rule-following and Institutions: An Evolutionary Perspective, in Mäki, U; B. Gustafsson and C. Knudsen (eds.) *Rationality, Institutions and Economic Methodology*, 171—200. London: Routledge.

Vanberg, Viktor (1994) *Rules and Choice in Economics*. London: Routledge.

Vanberg, Viktor (1994b) Cultural Evolution, Collective Learning, and Constitutional Design, in Reisman, David (ed.) *Economic Thought and Political Theory*, 171—204. Dordrecht: Kluwer Academic Publishers.

Vanberg, Viktor (1996) Institutional Evolution Within Constraints. *Journal of Institutional and Theoretical Economics*, 152, 690—6.

Vanberg, Viktor (1997) Institutional Evolution through Purposeful Selection: The Constitutional Economics of John R. Commons. *Constitutional Political Economy*, 8, 105—22.

Vanberg, Viktor and James M. Buchanan (1986) Organization Theory and Fiscal Economics: Society, State, and Public Debt. *Journal of Law, Economics, and Organization*, 2 (2), 215—27.

Vihanto, Martti (1993) Social Contract, Natural Law and Spontaneous Evolution: An Austrian Perspective. *Journal des Economistes et des Etudes Humaines*, 4 (1), 65—92.

Vihanto, Martti (1998) Using Psychology to Reinforce the Austrian Argument for Freedom: The Case of Loan Decisions. *Constitutional Political Economy*, 9 (4), 303—21).

Vromen, Jack (1995) *Economic Evolution. An Inquiry into the Foundations of New Institutional Economics*. London and New York: Routledge.

Waller, William T. Jr. (1988) The Concept of Habit in Economic Analysis. *Journal of Economic Issues*, XXII (1), 113—26.

Weber, Max (1946) *From Max Weber: Essays in Sociology*. Oxford: Gerth and Mills, trans.

Weber, Max (1947) *The Theory of Social and Economic Organization*. Oxford: Henderson and Parsens, trans.

Wernerfelt, Birger (1997) On the Nature and Scope of the Firm: An Adjustment Cost Theory. *Journal of Business*, 70, 489—514.

Whyte, W. F. (1955) *Money and Motivation. An Analysis of Incentives in Industry*. Westport: Greenwood Press.

Williams, George W. (1966) *Adaptation and Natural Selection*. Princeton NJ: Princeton University Press.

Williamson, O. E. (1971) The Vertical Integration of Production: Market Failure Considerations. *American Economic Review*, 61, 112—23.

Williamson, O. E. (1985) *The Economic Institutions of Capitalism*. New York: Free Press.

Williamson, O. E. (1991) Comparative Economic Organization: The Analysis of Discrete Structural Alternatives. *Administrative Science Quarterly*, 36, 269—96.

Witt, Ulrich (1986) Evolution and Stability of Cooperation without Enforceable Contracts. *Kyklos*, 39 (2), 245—66.

Witt, Ulrich (1991) Economics, Sociobiology, and Behavioral Psychology on Preferences. *Journal of Economic Psychology*, 12, 557–73.

Wolff, Rirgitta (1997) Constitutional Contracting and Corporate Constitution, in Picot, Arnold and Ekkehart Schlicht (eds) *Firms, Markets, and Contracts. Contributions to Neoinstitutional Economics*, 95—108. Heidelberg: Physica-Verlag.

Young, H. Peyton (1996) The Economics of Convention. *Journal of Economic Perspectives*, 10 (2), 105—122.

Zhou, Xueguang (1997) Organizational Decision Making as Rule Following, in Shapira, Zur (ed.) *Organizational Decision Making*, 158—88. New York: Cambridge University Press.

Zimbardo, P. and M. R. Leippe (1991) *The Psychology of Attitude Change and Social Influence*. Philadelphia: Temple University Press.