A SURVEY ON THE NATURE, REASONS FOR COMPLIANCE AND EMERGENCE OF SOCIAL NORMS

Gianluca Grimalda[*]

Introduction

The concepts of convention and norm have attracted the attention of many scholars from disparate areas of research, ranging from economists to philosophers. Arguably, the reason is to be found in that these concepts seem to bridge the gap between the sphere of individual action and that of social institutions, thus satisfying the tenets of the current dominating epistemological principle in all social sciences, reductionism. Not surprisingly, the definitions of conventions and of norms are all but definite; this is not due exclusively to some ambiguity intrinsic in the concepts themselves, but also to the growing application of the instruments of Game Theory, and in particular of the Evolutionary approach, to this field, which have brought out the necessity of more careful definitions, or even refinements, of these concepts.

The aim of the paper is then to survey some of the most important, to my view, contributions on the subject, hoping to tidy up this intricate matter. In section 2 it will be advanced a distinction between *internal* and *external* motive to action that will prove helpful throughout the analysis. Such a dichotomy will make it possible to divide the account of rules of behaviour in three categories, depending on the relation between these two basic prompts to action. The narrower concept is given by what I shall call *strictly-conceived conventions*, coinciding with Lewis's classical account of convention (section 3). We can then define *broadly-conceived conventions*, based on Sugden's early works on the subject (section 4) and, finally, *norms* (section 5).

In section 6 the problem of the compliance to norms will be examined, thus shifting the argument to the moral content of such rules. In section 7 a model of spontaneous emergence of rules of behaviour resting upon the instruments of Evolutionary Game Theory will be presented. The last section draws the conclusions of the arguments raised, and indicates further lines of analysis that will the subject of future research.

Internal and *external* motives to action

Before starting off I need to present what seems a widely accepted distinction in this literature, which will form the thread of the argument in the following analysis. As Harsanyi (1969) puts it, *people's behaviour can be largely explained in terms of two dominant interests: economic gain and social acceptance*. Likewise, Bicchieri (1990: 838) stresses that *a longstanding tradition in the social sciences contrasts instrumental rationality and social norms as alternative ways of explaining action.* The first terms of these dichotomies refer to concepts to which economists are more familiar: they enlighten the self-interested behaviour of rational agents, willing to choose the action suitable to maximise their given objective function (for instance profits in the case of firms, utility in the case of individuals). The second terms bring up a second relevant motive to action, which may be stated as a will to conform to the general norms reigning in a society, or to a search for *social status*, as stressed in psychological and sociological investigations (see for instance Coleman, 1990). The latter element clearly differs from the former in that it requires the observation of the behaviour of the other members of the society before choosing an action, thus possibly removing the possibility of an individualistic approach to the problem of rational choice.

The literature abounds with models of choice taking into account both motives to action: for instance Pettit (1990: 726) reduces the second motive to an *indirect* form of *self-interest*, whereby the agent contemplates the *esteem, affection, or pleasure* with which other members of society view her actions, and which can be added to the *direct* form of self-interest, of more direct economic significance and in principle measurable in monetary terms. The two types of interests, the *economic* and the *social*, make up the *overall interest* of the individual. Sugden (1998a) upholds a very similar view, in which the self-interested (or material) motive is weighed up with a quest to live up to the expectations of other agents, expressed in terms of their expected material payoffs. Margolis (1990) argues that an "optimal" balance between the two motives to action can be found by means of a properly "economic" calculus, taking into account the material and immaterial resources that each agent can freely transfer between the self-interested and the social goal. Furthermore, a "Darwinian" argument of selection *between* and *within* groups makes it possible to state in general terms such a principle of optimality, depicted by the maxim "neither selfish nor exploited" (Margolis, 1990: 824). Ben Ner and Putterman (1998) introduce a third motive to action, the *process-regarding* one, which adds to a *self-regarding* and an *other-regarding* one. This last addition should take into account not only the outcomes that are obtained, but also the ways in which those outcomes are reached, thus including a specific "moral" ingredient in the objective function. What all of these contributions have in common is the possibility of weighing up economic rewards and social rewards in this "extended" objective function, so that it is possible to apply the usual techniques of rational choice (selection of the best means to achieve given ends) with respect to such an objective function.

However, Elster (1990: 872) directly criticizes this view by arguing that when the economic and social motive are contrasted, then the difference between means and ends becomes blurred, as some social norms – possibly, all of them – are actually identifiable through the very means used to reach the desired outcomes[1]. Hence, he links the rationality of behaviour to the pursuing of self-interest, since this would be the only case in which such a distinction is neatly maintained. Other scholars put forward a similar argument, by claiming that these two motives to action are incommensurable, thus leading to the impossibility of the unity of practical reason (Copp, 1997). Finally, some authors adopt an intermediate stance, arguing that although a direct comparison is not always possible we may think of the social viewpoint as imposing some constraints to the individual self-interested choice (Rabin, 1995), or simply attaching an extra-value to the internal payoff (Sacco, 1997).

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

"Strictly-conceived" conventions

Co-ordination problems

In Lewis's classical study, still representing an obligatory reference point for all the contributions on the subject, the notion of convention only applies to a fairly limited variety of social interactions, namely the so-called *coordination problems.* These are defined as situations in which there exists a relevant coincidence of interests between the agents, and there exist two or more outcomes accruing equivalent, or very similar, payoffs to the agents.

Regarding the first point of the definition, when can we consider a situation as one of *predominant* coincidence of interests? Intuitively, following Schelling (1960, pp. 83-118, 291-303) we may think that all types of games can be ordered on a spectrum having at the two extremes games of pure coincidence of interests – where agents' payoffs, possibly after suitable linear rescaling, are equal in every square - and of pure conflict, namely zero-sum games. Accordingly, games with *predominant* coincidence of interests are fairly close to the former of these extremes, although they are internal to the continuum.

Consequently, *pure co-ordination games* are defined as games with perfect coincidence of interests having at least two Nash equilibria with equal payoffs, like in the following example:

C1	C2

R1	1,1	0,0
R2	0,0	1,1

The Battle of the Sexes is an example of a co-ordination problem where the coincidence of interests is still predominant, but is not perfect as the agents have different preferences over the outcomes.

A formal definition of a situation of predominant coincidence of interests can be worked out introducing the concept of co-ordination equilibrium. This is an outcome satisfying the following two conditions, for every agent:

a) one's strategy is a strict best reply to others' (strict *Nash* condition)[2];

b) given one's action, the actions played by her opponents maximize her payoff (*mutual benefit* condition).

Hence, the property of predominant coincidence of interests consists in that not only does each agent maximises her own utility, but, by doing this, she also maximises her opponents' payoffs given their strategies.

Summing up, on Lewis's account a co-ordination problem arises when at least two strict mutually beneficial Nash equilibria are present in a game. The "problem" lies in that, in spite of the presence of coincidence of interests, the presence of more than one of such mutually beneficial equilibria requires that some additional piece of information are added to the game in order to obtain co-ordination. This requires the analysis of the concept of expectation, upon which a more complete definition of convention will be put forward in section 3.3.

Expectations

Systems of expectations

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

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

a) others' preferences about the outcomes;

b) others' degree of rationality[4];

c) what you believe about the matters of fact that determine the likely effects of your alternative actions (Lewis, p. 27).

Undoubtedly the last element is the most controversial: in fact, it suggests that each agent will attempt to replicate others' reasoning to predict what their actions will be. This will generate a chain of mutual expectations, reaching an order of infinite level. Notice that this infinite-long chain of expectations exists only on a mere logical ground. They may be called a support to expectations, as only adding ancillary hypothesis of rationality to them can form proper expectations5.

Expedients to generate expectations

If the analysis conducted so far explains what is the *formal* structure of our expectations, it does not say anything about their *substantive* content. Clearly, some element external to the structure of the game is needed in order to orientate one's belief on others' behaviour. Lewis distinguishes three elements able to confer a substantive content to expectations.

The first is given by *agreements*. It is straightforward to notice that the co-ordination problem could easily be solved if the agents had the possibility to communicate before playing the game and reach an agreement about the final outcome. Being the possible equilibria mutually beneficial and equally worthy for the agents, one can be sure that the agents would not get stuck in a stalemate.

Another important means of co-ordination is given by the concept of *salience*. It is possible that a particular outcome in the set of the co-ordination equilibria *stands out from the rest by its uniqueness in some conspicuous respect* (Lewis, p. 35), as for instance number 1 in the set of natural numbers, or the centre of a city in the list of all possible meeting places. In fact, a large amount of experimental evidence has been gathered regarding how in certain situations agents found immediate co-ordination on a certain outcome without communicating (Schelling, 1960). In order to realise this form of spontaneous co-ordination agents must have some traits of cultural background in common.

Finally, another possibility for solving a co-ordination problem is just the previous occurrence of a successful co-ordination that forms a *precedent* for the future interactions. Suppose that today we have met in place A for mere luck. If tomorrow we face the identical problem of meeting, it seems clear that it is likely that both of us will head for A. In some way, A has now acquired a particular salience for the mere fact to be verified. The question of the stability of this mechanism to generate concordant expectations will be further analysed in the following sections, as many authors ground the foundation of the sense of morality of a community on this concept.

Common knowledge

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

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

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

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

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

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

And (2) applied to (4) implies:

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

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

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

A definition of convention

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

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

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

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

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

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

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

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

1) everyone conforms to R;
2) everyone expects everyone else to conform to R;
3) everyone has approximately the same preferences regarding all possible combinations of actions;
4) everyone prefers that everyone conform to R, on condition that at least all but one conform to R;
5) everyone would prefer that everyone conform to R', on condition that at least all but one conform to R',

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

Recalling the distinction set out in the introduction, and furthering some arguments put forward in section 5, we can immediately notice how in the case of strictly-conceived conventions there is no contraposition between internal and external motives to action, since social rationality and individual rationality clearly lead to the same outcomes.

"Broadly-conceived" conventions

Sugden's theory of moral conventionalism

As illustrated in the foregoing section, the basic features of a strictly-conceived convention are the presence of predominant coincidence of interests between the agents and the existence of more than one Nash equilibrium with equivalent rewards for the participants. However, Robert Sugden argues that between these two elements only the latter is actually necessary in order to define a notion of convention and build a comprehensive theory of morality upon it. Indeed, the former element is immaterial for that purpose, as will be clarified extensively in section 6, in the sense that the resilience of a certain regularity of behaviour can be guaranteed through the net of reciprocal expectations that is naturally developed among the members of a community.

This view enables Sugden to pool various types of regularities of behaviour into a wide class, virtually covering the whole class of interactions occurring in a society and not limiting itself to coordination problems. All of these rules share a "conventional" character, in that they are all arbitrary patterns of behaviour that emerge after a process of "selection" between the rules themselves.

In fact, the final goal of Sugden's work is to provide a comprehensive theory of the sense of morality, grounded on Hume's theory of moral conventionalism. His analysis may be deemed as *normative* in the peculiar sense that it describes people's current perception of morality, that is to say how they *think* they ought to behave, but not in the sense of supplying a series of normative prescriptions about how people *ought* to behave. Therefore, his analysis closely resembles one of psychology of morals, rather than dealing with the logic of moral propositions.

The notion of Evolutionarily Stable Equilibrium

In *The Economics of Rights, Co-operation and Welfare* Sugden defines a convention as any Evolutionarily Stable Equilibrium (ESS) in a game that contains two or more of such equilibria (Sugden, 1986, p. 32). The concept of ESS has been carried over from the field of biology to the literature of Evolutionary Game Theory to be applied to a situation in which there exists a large population of agents who are drawn at random at each instant of time and pairwise matched to play a certain game (Maynard Smith and Price, 1973; Maynard Smith, 1982). The agents are pre-programmed to play a given strategy, so that the population can be described in accordance to the percentage of players adopting each strategy. At the end of each repetition of the game, however, the average payoff gained by the whole population becomes known to the population, providing the agents with the opportunity to turn their strategy to a more profitable one (given the current distribution of the strategies that are played). However, the adjustment towards optimality is slow, as agents are supposed to be boundedly rational.

A first requirement for a notion of equilibrium in this context would require that agents did not need to change their strategies in any instant of time. This would indeed be suggestive of stability. However, it is apparent how this condition would be trivially satisfied in every situation where, by chance, all the agents played the same strategy and did not have any possibility "to imitate" any different strategy. The concept of ESS builds upon this idea by requiring that the candidate to the role of "equilibrium" in an evolutionary game must be "robust" to "invasions" of other strategies. In other words, assuming that for an arbitrarily small ε, the (1-ε) percent of the population play strategy A and ε "invaders" play B; then B cannot spread across the population over time.

Although the ESS is a static concept of equilibrium, it has a typically "evolutionary" flavour by means of such an "evolutionary" story that defines its intuitive meaning, which indeed will provide the ground for the account of the emergence of norms in section 7. Notwithstanding this, it can be shown that every ESS must be a strict Nash equilibrium in the stage game. This implies that, with respect to Lewis's definition of co-ordination equilibria, the requirement of individual rationality – i.e. the Nash equilibrium condition- still holds, while that of a mutually beneficial equilibrium is lacking. This makes it possible to define another class of "conventions", which I shall call "broadly-conceived" to emphasize that they apply to a wider area than those emerging from co-ordination problems solely. In what follows I will illustrate two types of conventions that are strict Nash equilibria without the property of mutual benefit, aiming to enlighten in which sense a moral meaning can be attached to such interactions.

Conventions of property

The interaction in the state of nature

The claim that property rights are the result of a process of convergence to a convention when individuals interact in a state of nature is one of the most appealing results of Hume's theory of morality. The starting point is Hobbes's classical model of a state of nature. The results of his characterisation are well known: when there is a situation of rough equality between men, and the main goal of people is conceived to be their own survival, thus seeking to ensure every means and to undertake every action to support this end, we end up in a state of war of everyone against everyone. All are worse off by this result, since resources and means are allocated to the aim of defending themselves in front of others' possible attacks instead of being employed in productive activities. Hobbes's response to such a conflict situation was the appeal to an authority to which individuals would have transferred part of their power in order to ensure the enforcement of peace and agreements.

Hume's answer was different: the evolution of interactions between people in a state of nature would *spontaneously* lead to the emergence of rules that assign property rights to individuals, with no need of an authority. Such an allocation of goods is conventional, in the sense that the equilibrium reached does not respond to any external criterion of efficiency, fairness, or justice, as it occurs in the contractarian case.

The hawk-dove game

Sugden employs various models of games to depict the typical situation of fight between people for the property of goods. I will focus on the simplest of them, the hawk-dove game. The matrix of payoffs is reported below:

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

It is assumed that when people desire the same thing, which nevertheless they cannot both enjoy, they may either incur in a fight, generating a mutually destructive outcome, when both claim the good (Hawk-Hawk), or in a peaceful sharing of the good (Dove, Dove), or finally in the attribution of the thing to one of the two, when only an agent claims the good and the other renounces.

What equilibrium will be obtained? We know that there are just two strict Nash equilibria in this game, given by the two outcomes where one agent plays Hawk and the other Dove. Therefore, the "convention" which will emerge will be one in which the property of the good is assigned to one of the two agents. In section 7.3 we shall illustrate how such an outcome can actually emerge in a dynamical context.

Conventions of reciprocity

The typical example of a reciprocity game is given by the well-known Prisoner Dilemma (PD). If we take the static game there is clearly no possibility for a convention to emerge, given the existence of a unique Nash equilibrium. However, many commentators have argued for the possibility of defining a convention even in this case (Sugden, 1986: ch. 6; Hardin, 1988). If we consider the super-game made up by the repetition over a possibly infinite period of the one-shot PD, with a positive probability of the super-game suddenly ending after a finite number of periods, an infinitely large number of possible ESS exist. Picking up for simplicity just two of these, that is the well-known tit-for-tat 6and the strategy prescribing to defect at every game - the so called nasty strategy-, we end up with the following matrix of payoffs gained in the super-game:

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

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

This account may be criticized on the grounds that the couple of ESS here considered falls short of one of the constitutive properties of a convention, namely the “comparability” of the competing outcomes (see section 3.2). In fact, the equilibrium given by (T,T) will give a remarkably higher reward to the agents compared to the equilibrium (N,N), thus casting some doubts over the possibility to consider each equilibrium as a proper alternative to the other. Even though these equilibria may formally satisfy Sugden’s requirement, it is at least dubitable that the emerging pattern of behaviour may be considered “a matter of convention”.

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

Norms

So far we have analysed conventions sustained by some sense of self-interest. In the first class, co-ordination games, there is a proper coincidence of interest, given the mutual benefit brought about by the emergence of the convention. In the other two classes of games, property and reciprocity games, the characteristic of mutual benefit is lost, although the conventions that are established are Nash equilibria. In this section we want to see how regularities of behaviour detrimental in terms of self-interest can nonetheless emerge. In section 5.1 I shall present a workable definition of norm, while in section 5.2 I will analyse in greater detail the notion of normative expectations, who takes on a crucial role in most of the theories on the subject.

A definition

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

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

(1) nearly everyone conforms to R;

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

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

Such a definition is clearly modelled on that of Lewis for conventions, and it differs from it for some important respects.

Arguably, the most relevant difference lies in that there is no requirement about individual self-interest: in fact, the requirement of approval and disapproval to the norm substitutes the condition based on individual preferences and on reciprocal expectations. Of course, if the norm satisfies the individual's self-interest, then approval in case of conformity and disapproval in case of deviance may be expected to emerge, especially in situations of coincidence of interests between individuals. However, relying on a notion of approval and disapproval as the main support of a norm opens the way to regularities of behaviour that may go against the direct self-interest of individuals but may be sustained because of an external motive to action. In other words, the sense of obligation to follow a norm may offset one's self-interest, and suffice to grant the compliance to it.

In most cases the norm will fulfil some notion of public interest: this will provide the agents of a community with a valid external reason to comply with the norm. Accordingly, many authors focus on the so-called norms of cooperation, typically emerging in a PD-like situation. For instance, Pettit deals with patterns of behaviour satisfying what he calls the interaction assumption: “among the options available to any agent in the sort of situation involved nearly everyone is better off if everyone else takes one particular option than if everyone else rejects it: the option in question is, in that sense, a collectively beneficial one” (Pettit, 1990: 743). Such a condition does not require that the action prescribed by the norm fulfils anyone’s self-interest, but that it is reciprocally beneficial for everyone, thus comprising Pareto dominant allocations. This argument may provide an explanation of the emergence of cooperation even in static PD, not only in repeated interactions (see section 4.4 and also 5.2). Besides, Bicchieri (1990) explicitly focuses on repeated PD-dilemmas where agents can assume some predetermined types of strategies.

However, norms sustained by external motives are not limited to collectively beneficial outcomes. For instance, the so-called rules of courtesy lack the possibility of singling out a unique Pareto-efficient outcome, thus requiring to some agents to be in a position to yield to others. An example of such regularities is given by the so called rules of courtesy: in many cases someone may opt to follow a behaviour with the only aim to be kind to someone else, even if this is contrary to the agent’s direct self-interest. For example, someone may decide to renounce to her seat on the bus to offer it to an old woman. Either, take the example of many tacit rules of the highway code: someone could decide to give way to a car even if she is on the main road and the highway code of the road allows me to go straight ahead. On a larger scale, we may envisage social situations in which the sense of public interest requires to some groups of the community to sacrifice their welfare for the benefit of other groups.

Sometimes norms may turn out not to be beneficial to anyone involved in the interaction, like in norms of revenge: in this case, the most plausible explanation advanced by students of the subject is that they are the result either of a psychological disposition that may be useful in situations different from the mutually destructive one, or that it was necessary in some past phases of human evolution (see Elster, 1990: 885; also Frank, 1988). In neither case, according to those authors they can be justified by some form of public interest or individual rationality.

Therefore, norms cover an area much wider than conventions: conventions can be considered norms, in a sense that will be further clarified in section 6, but there exist norms that cannot be thought of as conventions, since they lack the condition of individual self-interest. Most scholars view norms as peculiar institutions that reinforce or overlap with other patterns of behaviour, like customs, laws or social standards. Thus, for instance a certain pattern of behaviour may at the same time be a norm and a law8.

That the crucial character of a norm is the sense of obligation implicit in it can also be seen by the fact that a norm must also be socially enforced, that is publicly sustained by people not directly involved in the interaction. On Lewis’s account, a failure to conform by some agent tends to evoke unfavourable responses from all others agents, even those not directly involved in the instance (Lewis, p. 99). According to Hume, there are two reasons for this. The first is immediate: we typically feel a sentiment of sympathy toward others' experiencing a certain situation, especially when we are in a close relation to those people-for instance fir a parental or a physical relation. The second is mediate: we could think that in the near future we could be involved in the same kind of situation of which now we are only bystanders. Therefore, we may be led to condemn someone breaching a norm for the fear to meet her in our next interactions, or for the knowledge that the imitation on a wide scale of such a deviant behaviour by other people could end up to be detrimental for my interests.

Finally, another relevant difference between norms and conventions lies in that we do not need to have a viable alternative to the rule currently followed by the population. Therefore, the idea of arbitrariness implicit in conventions is not required for norms.

The resentment hypothesis and the role of normative expectations

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

"What reward is most proper for promoting the practise of truth, justice and humanity? The confidence, esteem and love of those we live with. Humanity does not desire to be great, but to be beloved." (Smith, 1759/1982, p. 166). "We are pleased to think that we have rendered ourselves the natural objects of approbation, though no approbation should ever actually be bestowed upon us: and we are mortified to reflect that we have justly merited the blame of those we live with, though that sentiment should never actually be exerted against us"(Smith, 1759/1982, p. 116).

This simple but fundamental assumption about human psychology is called by Sugden the resentment hypothesis (Sugden, 1998a, p. 16). It takes on a very important role in a theory blending internal and external motives to action since it permits to found the external motivation in the individual system of deliberation, thus making it directly comparable to the internal spur to action. In other words, by means of the resentment hypothesis the external motivation is "internalised" in one's system of choice, thus making the "overall interests function" (see section 2) a sound construction[9]. Besides, by grounding the function and emergence of norms on individual motivations, the reductionist approach is somewhat safe, since a theory that overlooked this point would run the risk to explain norms in terms of the existence of other norms (Sugden, 1998a, p.4).

Expectations come to be considered "normative" since they provide us with a strong sense of commitment to the pursuing of the rule generally followed in the community, conferring the character of obligation typical of norms (see the previous section). When a rule is established, each agent understands that its breaching would trigger a sense of resentment in other members of the community and as a consequence a sense of guilt in ourselves, thus urging us to refrain from flouting it (at least until the possibly contrasting self-interest becomes too strong).

However, expectations take on a cognitive aspect too. Indeed, agents perceive the norm by means of the set of expectations that all of the other agents of the community -the direct opponents of the interaction and the bystanders not directly involved in it- address to her. This introduces a further reason for which agents may want to abide by the rules of a community. For the norm may indicate what the public interest of a community is, thus inducing an individual to follow it because of her internal commitment to act in accordance to the general public interest, rather than to the resentment hypothesis (Pettit, 1990: 731). This element may be relevant in that sometimes the public interest is in contrast with the reigning norms, in those situations in which a norm is patently "wrong" or inefficient. This argument brings on the question of the stability of a norm, and of its change – or revision – over time (see Ulmann-Margalit, 1990), which will be further analysed in the next sections.

Also, many scholars doubt that the account of cooperation in a Prisoner Dilemma-like situation grounded on the idea of the adoption of a tit-for-tat strategy is actually successful (see section 4.2). It has been pointed out that, especially in the many-players PD, the interaction may be such that the defection of one single agent may cause so negligible costs that the punishment from the rest of the agents may be economically inefficient. In these situations the tit-for-tat may not be a viable self-enforcing strategy to bring about an outcome of reciprocal cooperation (Pettit, 1989a: 341-344). Therefore, so the argument goes, an explanation based on normative expectations gains credibility in these contexts, since the costs implied by the mere observation of others' behaviour and the consequent sentiment of commendation or disapproval are virtually nil[10].

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

1) Interaction assumption, referring to the collectively beneficial character of the norm (see the previous section);
2) Publicity assumption, stressing the necessity that the behaviour of the agents involved be observable by the others involved;
3) Perception assumption, referring to the possibility for the agents to clearly make out whether everyone's action was for or against the benefit of the community;
4) Sanction assumption, that underlines how the enforcement of a norm can be brought about by the attitude to encourage the obedience of the norm and discouraging its transgression embedded in agents' dispositions;
5) Motivation assumption, equivalent with the resentment hypothesis.

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

The problem of compliance with norms and the question of their moral content

So far we have analysed the concepts of convention and norm without stressing in great detail the reasons to comply with them, especially in the case of norms sustained principally because of external reasons. Strictly connected to this question is that regarding the moral content that can be attributed to those patterns of behaviour. This is the subject of the present section.

Presumptive reasons for conformity to norms

The point of departure in accounting for the reasons to comply with conventions and norms is the same, and consists of the so called presumptive reasons to conform to a convention.

Strictly-conceived conventions

Mutually beneficial Nash equilibria are sustained firstly because of self-interest. To be sure, this gives a straightforward reason to comply with them. But the contemporary presence of the element of the reciprocal benefit in the pursuing of one's self interest also attributes a specific normative and moral character to these regularities.

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

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

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

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

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

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

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

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

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

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

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

Although the adverb “especially” seems to imply that in some occasion it might be true that one follows only her external motives to action, thus satisfying others’ preferences without considering her own, Lewis seems to consider this occurrence a rather exceptional event. In the following passage he clearly claims that the two presumptive reasons must both be present in order to support a convention: “for any action conforming to any convention, we would recognise these two (probable and presumptive) reasons why it ought to be done” (Lewis, p. 98). In this sentence, the adverb “reasonably” refers to the fact that by following the convention I am answering to my preferences, as I will clarify in section 6.2.1. The others have reason to believe that I will conform as long as they know that it is in my interest to conform.

This point is further clarified by Lewis when he deals with socially enforced norms. On Lewis’s account, the sentiment of disapproval I invoke in people part of my society comprises both a feeling of resentment for not having answered others’ preferences, and a feeling of surprise to have acted contrary to my own preferences. “So if they see me fail to conform, not only have I gone against their expectations; they will be probably be in a position to infer that I have knowingly acted contrary to my own preferences and their reasonable expectations. They will be surprised, and they will tend to explain my conduct discreditably (Lewis, p. 99).

It is for the presence of these two kinds of “ought” derived from the pair of presumptive reasons for conformity that the normative character of this type of convention becomes apparent. It is indeed straightforward to show that the three conditions set out in section 5.1 in order to define a norm are satisfied: of course they commence a regularity of behaviour (condition 1), which rises commendation on conformity and censure on deviance (condition 2) from all the members of the community, and every agent can understand that these sentiments elicited in members of the community are reasonable, thus strengthening the compliance with the norm (condition 3).

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

Other kind of norms

The two presumptive reasons are both present in the narrower concept of conventions. As far as the equilibrium is mutually beneficial, I both have an interest in following the rule and others have an interest that I follow the rule, thus giving me a further motive to follow the rule. The problem with broadly-conceived conventions is that they are not always mutual benefit equilibria. On Sugden's account, however, a moral reason to abide by broadly-conceived conventions and, generically, norms can be grounded on the concept of normative expectations, and in the possibility of considering the two presumptive reasons set out above independently from each other. In fact, the knowledge of others' expectations takes on the status of a moral commitment in virtue of the idea of normative expectations.

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

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

Some critical remarks on the concept of normative expectation

On Sugden's account, for normative expectations to act as a guide to action it is sufficient that a regularity of behaviour has established. Whenever the agents know that, then they will form the expectation that agents will behave in a way coherent to such a regularity. Notice the different usage of the adjective "reasonable": while in the case of strictly-conceived conventions this both implied an internal and an external motivation (section 6.1.1), in the case of more general norms one can find "reasonable" whatever type of behaviour is eradicated in the habits of a community and has acquired sufficient regularity, notwithstanding the relationship to someone's self-interest (section 6.1.2).

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

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

When is an expectation reasonable? Empirical and causal expectations

The argument in this section is related to the use of the term "reasonable" that allows Sugden to introduce the concept of well-grounded empirical expectations. As he refers to Lewis's work to introduce this concept, I will do the same. Throughout his book Lewis uses the expression "to have reason to X", where X can be either the verb "to desire", or "to expect", or "to believe", as a part of a peculiar inference, which has the following structure:

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

Further, translating this inference to a higher order,

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

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

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

Hence, I seem we can conclude that on Lewis's account both kind of expectations, causal and empirical, must be present in order to make up a firm reasonable expectation. We can infer that it might be the case that people form their expectations on the ground of a precedent, thus forming empirical expectations only, but this would give a much less stable basis for the concordance of expectations. This remark clearly does not undermine the whole of Sugden's point, but stresses how the claim that all of Lewis's accounts of reasonable expectations are grounded on an empirical basis seems to miss some important aspects of his accounts of norms.

Conformative behaviour and normative expectations

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

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

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

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

Conclusion: I conform

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

The problem of this inference is that it explicitly relies on the rule itself, creating a circularity in the definition (the existence of the rule shapes my expectations, but actually it is a system of concordant mutual expectation which should give rise to a rule). However, according to Sugden, we actually do not need the concept of rule in this inference, but only the occurrence of a precedent. In other words, the precedent acts as a basis for common knowledge about our expectations. Therefore, we can restate the previous inference eliminating any explicit reference to a rule:

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

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

Conclusion: I conform

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

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

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

Conclusion: I expect that you expect that I will conform

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

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

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

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

The most relevant problem with Sugden's theory is that expectations seem to be a set of stable parameters easily recognisable by each agents. Take the case of a repeated prisoner dilemma, and suppose that the current rule prescribes that one agent defects and the other cooperates. Formally, normative expectations would suffice to sustain such an outcome, since the gain that the co-operator would get if he defected could be outweighed by the resentment for not having lived up to others' expectations within her overall interests function. But what about the formation of their expectations? The defector knows that the rule imposes a heavy cost on C, and she may fear that one day C will begin to be fed up of only serving D's own purposes. Her expectation about C's conformity cannot be very strong, if she takes into account the opponent's direct interests. Further, if C is sufficiently rational, she will anticipate D's doubts about her behaviour. Therefore, she may perceive that D's expectations about her is not so strong as she may have believed in the past. When forming her second order expectations, she may take this into account, and believe that her expectation that D expects her to conform is not so firm. Then D should allow for C's perplexity when forming his third order expectations, and so on. The point is that nearly all of the outcomes of the game may be considered equilibria if sustained by an opportune set of expectations. And to point to normative expectations in order to elicit compliance to a conventions is tantamount to saying that in a society dominated by slavery slaves conform to the rule because of the sense of resentment they would elicit in their masters. Expectations based on empirical reasons but not grounded on individual interests would fail to be a basis for common knowledge: with rational agents the expectations would not extend much far beyond the first levels of mutual expectations, thus undermining the stability of the rule.

Claiming that expectations make up all the normative content of conventions is in my view far-fetched: surely they help to strengthen the commitment to a norm whenever this is established, but they do not exhaust its explanation. I think something more is required in order to account for the whole cognitive system beneath a set of concordant mutual expectations and the sense of justice people experience, and this further element should come from the idea of public interest that people perceive in a rule, of which the concern for others' *individual* interests involved in the interaction, transmitted through others' expectations, is a constitutive part[12]. Modelling this element will form the subject of future research.

A second reductionist approach: the dynamic process of convergence towards a convention

The fallacies intrinsic in an equilibrium analysis

The main problem with the analysis set out in the previous section is, to my view, that an equilibrium analysis, namely a static concept like that of the ESS supported by the notion of normative expectations, is not sufficient to account for the whole character of a rule of behaviour. Indeed, claiming that a norm is sustained by the sensation of social disapproval that the agent would perceive when flouting it – or even the associated material sanctions in case the norm is enforced by a legal system – reduces all of the reasoning again to a *cost-benefit* argument. This is, however of little help in explaining why norms are there, how they emerged and why they persist. In fact, nearly every outcome of a game could be sustained by means of a "normative expectations" story. Although this is functional to Robert Sugden's theory of conventionalism, to be sure it leaves many problems unanswered, as I have sought to show in the previous section.

In Bicchieri's words (1990, p. 839), *to say that one conforms because of the negative sanctions involved in nonconformity does not distinguish norm-abiding behaviour from an obsession, in which one feels an inner constraint to repeat the same action in order to quiet some "bad" thought*; or, to put it in other words, actions complying with norms may be seen as a categorical imperative[13]; but then the reductionist endeavour of grounding social institutions on individual deliberation would be frustrated.

The other reductionist approach of justifying norms on the grounds of their *collective* rationality is, according to Bicchieri, bound to fail as well. In this view a norm would be justified for its efficiency in pursuing a certain end, like for instance social welfare. But this would lead to a *post hoc ergo propter hoc* fallacy, since the mere presence of a social norm does not justify inferring that it is there to accomplish a social function, and indeed in many cases the contrary may be upheld (Bicchieri, 1990: 838).

One possible way of escape from these criticisms is provided by the Evolutionary approach to games. Not only does it permit to analyse the equilibrium property of norms, but also it enables to focus on the process of their formation, emergence and persistence. When the concept of ESS was introduced (section 4.2) I already emphasized how this conveyed an evolutionary flavour. Now this is explicitly stated and made the central aspect of the story. Therefore, the dynamic account of norms differs from the static one in that it emphasizes the *how* norms emerge instead of the reasons *why* they did, thus providing a different account of the concept.

The emergence of a strictly-conceived convention

To give an account of how the Evolutionary approach to norms formation works, let us come back to Sugden's works and let us consider the kernel of the notions of norm, that of strictly-conceived conventions14. Let us consider a particular type of this category, the cross-roads game, a symmetrical game that represents a situation of partial coincidence of interests. Let us suppose to start from a point of absence of co-ordination among the individuals of the population (recall the setting of evolutionary games set out in section 3.2). The pure strategy Nash equilibria are mutually beneficial, thus satisfying the requirements of strictly-conceived conventions:

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

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

Conversely, if the recognition of some asymmetries in the labelling of the players makes the game an asymmetrical one, it is possible to reach equilibria in which agents play different strategies. The Nash Equilibria (Maintain Speed, Slow Down) and (Slow Down, Maintain Speed) are showed to be ESS, while the previous equilibrium in mixed strategy fails now to be stable. The basic idea is the following: let us begin from the situation in which each player slows down with probability 0.8; then, suppose that a percentage ε of the agents perceive an asymmetry of whatever kind in the game, say they start thinking that players coming from the right give way to other players. Therefore, they believe that it is convenient for them to maintain speed when coming from the left and to slow down otherwise. If we suppose that the probability of coming from either left or right is the same, such a belief turns out to give them a higher payoff. In fact, when two people of this group of “smart” agents meet each other, they gain a payoff higher than average, that is large enough to compensate the loss when they meet “dumb” players. Nevertheless, as dumb players perceive that people maintain speed with higher probability; they are compelled to slow down more frequently, or even each of the time. After some adjustment, the smart players successfully apply the convention within their group, while dumb players gain the same payoff as before. But the situation is likely to evolve. As soon as larger shares of dumb players recognise that the group of smart players is more successful, they will be willing to shift to the convention. Since the equilibria are mutual beneficial, the group of smart players do not have any reason to prevent them from doing so.

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

Emergence of broadly-conceived conventions

Let us now consider the Hawk-Dove game already introduced in section 3.3. Following the tassonomy offered by Weibull for symmetric games (Weibull, p. 40), this game is formally equivalent to a co-ordination game, thus implying the same type of equilibrium and the same dynamics of convergence towards them. Similarly to the previous game, the equilibrium of war of everyone against everyone in the state of nature is depicted by the equilibrium in mixed strategy when the game is played in the asymmetrical form, not instead by the outcome obtained when everyone is aggressive (H,H), which, would probably give the best representation of the state of war in the state of nature.

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

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

Some critical remarks

The role of asymmetries

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

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

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

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

b) if the game is not strictly symmetrical even from the formal point of view, then it is possible that the structure of the payoffs itself could generate a difference in the behaviour of the agents: for example, if the outcome of (maintain speed, maintain speed) gives a slightly better outcome to

Player 1 than to Player 2, then the agents who enter the games as Player 1 will have a small incentive to maintain speed with higher probability. When this feature will be recognised by agents who enter the game as Player 2, then a convention in which Player 1 maintains speed and Player 2 slows down is likely to arise;

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

Single and multiple focal points

My understanding of Sugden's argument is as follows: the dynamics of the process leading to the universal adoption of a convention rests on the possibility that agents recognise some sort of asymmetry capable of labelling in some way their participation to the game. This is clearly possible when an asymmetry stands out as an unique focal point, since this will act as a basis for common knowledge helping members of the population to generate a system of concordant mutual expectations. Nevertheless, it is not possible to rule out the possibility that the "set" of possible asymmetries in the game is multiple. In this case, Sugden relies on a sort of process of trial and error for which the population eventually succeeds in "converging" to recognize a single relevant asymmetry.

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

However, when the context of the game does not present a prominent asymmetry, the situation is much more problematic. The peculiar tools of Evolutionary Game Theory actually *assume* that the group of "mutant" agents adopt the same rule of behaviour. In fact, the assumption that mutations occur once at time is rather simplifying: *evolutionary stability is a robustness test against a single mutation at a time. In other words, it is as if mutations are rare in the sense that the population has time to adjust back to status quo before the next mutation occur* (Weibull, p. 34). This hypothesis could seem justifiable in a biological context, but it appears rather restrictive in a social context, where there are plenty of asymmetries capable of attracting the attention of the agents. In fact, the agents of the mutant group must experiment a problem of co-ordination analogous, if not worse, to the original one, because of the increased number of available alternatives. Thus the original problem seems simply shifted backwards, to the problem of choosing one of the multiple asymmetries that are likely to generate concordant expectations[15]. Of course, this may happen by chance; but the period of time needed to obtain such an homogeneous mutation may well be infinitely long, since its probability is rather small.

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

This problem is far more relevant if we introduce some changes in the formal structure of Evolutionary Game Theory to “adapt” it to the context of social interactions. Rather than thinking that a mutation in the general rule of behaviour by a restricted part of the population happens by chance, we could, more realistically, assume that this change is *voluntarily* brought about by a minority of “smart” agents, capable of reaching a certain degree of understanding of the complex of the interaction. In this way, the periodic mutations would not be a random process, but the result of a conscious evaluation by “enlightened” agents.

Now, in these circumstances the smart agents will recognise that every change in their behaviour is detrimental for them, unless all of them happen to identify the same asymmetry in the game, an event which has a too small probability to be considered. The result would be that no more mutations in behaviour would be brought about, thus obstructing from the beginning the evolutionary process that should lead to the general adoption of a convention.

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

This process would follow the same dynamics as that depicted by Sugden, with the peculiar difference that the basis for the common knowledge is now given by an agreement rather than by the prominence of an asymmetry, thus offering a sounder basis as a device to generate concordant and mutual expectations and fostering a quicker convergence toward the general adoption of a convention.

Possession in property and reciprocity conventions

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

Alternatively, we could think that there is a somehow universally shared *rule* to assign possession of things, that does not bring about any situation of conflict, upon which the property games can be sorted out. But what could such a rule consist of? Could it be represented by a concept of *geographical proximity* toward the thing object of the conflict, as many examples of Sugden's analysis could let us think of? But in this case a conflict for the possession of the best and richest area of the territory is likely to arise. Perhaps we could think that a somewhat *impartial* rule, that everyone could accept with no controversy, could be adopted as a device to solve this problem. But this would be antithetical to Sugden's approach: impartiality is in itself a moral concept, thus it should be the outcome of a process of evolution, not its starting premise.

The same sort of argument applies to the reciprocity conventions as set out in section 4.4. In that case a convention was seen as one of the possible tit-for-tat-like strategies with different period of punishment. In this case it is even harder to perceive a remarkable asymmetry capable of triggering the process of convergence to a stable outcome.

Conclusions

The main attempt of this work has been to survey the main contributions on the constitution and emergence of rules of behaviour in a society, drawing on the distinction between internal and external motives to action. These range from strictly-conceived conventions to norms depending on the weight assigned by each individual to self-regarding and external motives to action. This is to my view a very promising approach, which I shall try to formalise in future research drawing on the instruments of psychological games, which have proved a helpful instrument to give account of the presence of internal and external motives to action in an individual's system of choice (Geanakoplos et al, 1989).

However, I have also tried to show how an account of norms based exclusively on the idea of normative expectations may lead to relevant problem of stability in that the cognitive structure required to generate common knowledge of the compliance to the norm cannot be grounded on precedent only. I have argued that a greater attention to the interests of the agents involved, or to an idea of public interest that they may want to pursue, is indeed necessary in order to overcome this problem.

Furthermore, I have illustrated the model of emergence of rules grounded on the evolutionary account, as proposed by Robert Sugden. I have sought to enlighten, along with its merits, some of its shortcomings. In fact, it seems prone to a criticism again connected to the problem of the common cognition shared by the players in acknowledging prominent asymmetries in the games. Also, I believe that the theory may gain a good deal of insight by allowing for the possibility that small groups of agents may agree to coordinate their action on a given pattern of behaviour, which then would spread to the whole population by means of the evolutionary forces. In other words, it seems to me that integrating an evolutionary account with one based on the notion of agreement, possibly circumscribed to a limited group of agents, may considerably strengthen the power of the theory. This conviction is enhanced when one considers that a process of emergence of norms entirely based on evolutionary forces could be rather slow and burdensome for the population, and lead to inefficient outcomes.

Finally, theories of norms and institutions based on this broad set of motivations generally adopt an "equilibrium" analysis, thus overlooking the problems regarding their stability. This is reflected in the fact that, to the best of my knowledge, the evolutionary account of the emergence of norms has been applied only to standard games where agents' objective functions are "well-behaved", and not to psychological games, where instead the inclusion of expectations widens of various degrees the dimensionality of the objective function. Indeed, the problem of convergence when expectations are incorporated in one's objective function opens technical and substantive problems, but it seems a necessary task to undertake, since in a static context the notion of normative expectations make every outcome in a given interaction justified, even those clearly implausible. The analysis of the stability and of the possibility of convergence towards a norm under these condition, technically, to a psychological Nash equilibrium, will be the subject of the second paper of this series.

References:

Anderlini, Luca and Antonella Ianni: *Path Dependance and Learning from Neighbours*, Games and Economic Behaviour 13: 141-77

Axelrod, R.: *The Evolution of Cooperation*, New York: Basic, 1984

Ben Ner, Avner and Louis Putternam: *Values and institutions in economic analysis*, in Avner Ben-Ner and Louis Putterman (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 3-69, 1998

Bicchieri, Cristina: *Norms of Cooperation*, in: Ethics 100 (July 1990), pp. 838-861

Binmore, Ken*: Game Theory and the Social Contract, Volume 1: Playing Fair*, Cambridge, Mass: MIT Press, 1993

Buchanan, James: *The Limits of Liberty*, Chicago: University of Chicago Press, 1975

Coleman, S.J.: *Foundations of Social Theory*, Cambridge, Mass: Harvard University Press, 1990

Copp, David: *The Ring of Gyges: Overridingness and the Unity of Reason*, in: Social Philosophy and Policy, Cambridge University Press, Vol. 14 N.1, 1997

Elster, Jon; *The Cement of Society*, Cambridge: Cambridge University Press, 1989

Elster, Jon: *Norms of Revenge*, in: Ethics 100 (July 1990), pp. 862-885

Frank, Robert: *Passions within Reason*, New York: W.W. Norton & C., 1988

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

Geanakoplos, John, Pearce David, and Stacchetti Ennio: *Psychological Games and Sequential Rationality*, in: Games and Economic Behavior 1, 60-79 (1989)

Hardin, Russell: *Morality within the Limits of Reason*, Chicago: Chicago University Press, 1988

Harsanyi, John, *Rational Choice Models of Behaviour versus Functionalist and Conformist Theories*, in: World Politics, 22 (1969), pp. 513-38

von Hayek, F.: *Law, Legislation and Liberty*, vol. 1 *Rules and Order*, London: Routledge &Kegan Paul, 1973

Hume, David: *A Treatise of Human Nature*, Oxford: Clarendon Press, 1740 (reprinted 1978)

Lewis, David: *Convention: A Philosophical study*, Cambridge, Massachusetts: Harvard University Press, 1969

Mainard Smith, J. and G.R. Price: *The Logic of Animal Conflict*, in. Nature, 246: 15-18, 1973

Mainard Smith, J.: *Evolution and the Theory of Games*, Cambridge: Cambridge University Press, 1982

Margolis, Howard: *Equilibrium Norms*, in: Ethics, 100 (July 1990), pp, 821-837

Pettit, Philip: *Free Riding and Foul Dealing*, Australian Journal of Philosophy 67, 1989a

Pettit, Philip and Robert Sugden: *The Backward Induction Paradox*, Journal of Philosophy 86: 169-83, 1989b

Pettit, Philip: *Virtus Normativa*, in: Ethics 100 (July 1990), pp. 725-755

Rabin, Matthew: *Moral Preferences, Moral Constraints, and Self-Serving Bias*, Working Paper, Department of Economics, University of California, Berkeley, August 1995

Sacco, Pier Luigi: *On the dynamics of social norms*, in: Cristina Bicchieri, Richard Jeffrey and Brian Skyrms (eds.): *The dynamics of norms* Cambridge: Cambridge University Press, Ch. 3, 1997

Sacconi, Lorenzo: *Eduzione vs. evoluzione nella selezione dell'equilibrio: un'alternativa all'analisi dell'insorgenza dell'ordine in Hayek*, Quaderni di Storia dell'Economia, IV/1986/3, pp. 157-201

Schelling, Thomas C.: *Strategy of Conflict*, Cambridge, Mass.: Harvard University Press, 1960

Skyrms, Brian: *Chaos and the explanatory significance of equilibrium: Strange attractors in evolutionary game dynamics*, in: Cristina Bicchieri, Richard Jeffrey and Brian Skyrms (eds.): *The dynamics of norms* Cambridge: Cambridge University Press, Ch. 10, 1997

Smith, Adam: *The Theory of Moral Sentiments*, Oxford: Clarendon Press, 1759 (reprinted: 1976)

Sugden, Robert: *The Economics of Welfare, Rights and Co-operation*, Oxford: Basil Backwell, 1986

Sugden, Robert: *Contractarianism and Norms*, Ethics, 100 (July 1990), pp, 768-786

Sugden, Robert: *The motivating power of expectations*, mimeo, 1998a

Sugden, Robert:: *Normative expectations: the simultaneous evolution of institutions and norms*, in Avner Ben-Ner and Louis Putterman (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 73-100, 1998b

Taylor, Michael *The Possibility of Cooperation*, Cambridge: Cambridge University Press, 1987

Ulmann-Margalit, Edna: *The Emergence of Norms*, Oxford: Oxford University Press, 1977

Ulmann-Margalit, Edna: *Revision of Norms* Ethics, 100 (July 1990), pp, 756-767

Weibull, Jorgen W.: *Evolutionary Game Theory*, Cambridge, Mass: The MIT Press, 1995

Young, Peyton H.: *The Evolution of Conventions*, Econometric, 61: 57-84, 1993

Young, Peyton H.: *Individual Strategy and Social Structure*, Princeton: Princeton University Press, 1998

Notes

[*] I am grateful for helpful comments received from Alan Hamlin and Lorenzo Sacconi. I also wish to thank the financial support granted by MURST.

[1] This argument may be contrasted by the consideration that the same outcome reached though different means can be actually split into a set of different outcomes if the agent attributes intrinsic value to the means used to reach the outcome. Of course this same argument can be applied to Ben Ner and Putterman's approach.

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

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

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

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

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

[7] It should be pointed out that between the condition of approval of compliance and disapproval of deviance, the crucial one is the second. In fact, if we disapprove of someone's not doing an action and do not disapprove of her doing the action, this is tantamount to approving of the action. Conversely, we can approve of someone's performing an action and not disapproving of her not performing it, but this situation falls in category of super-erogatory virtues, which should not be considered norms (Pettit, p. 730).

[8] Elster (1989) undertakes quite a different route in accounting for norms. Not only does he seem to rule out the possibility of overlapping between these various classes of pattern of behaviour, but also he draws a difference between *social* and *moral* norms. He defines the first as *"nonconsequentialist obligations and interdictions, from which permissions can be derived"*, while the latter varies according to the moral theory taken as a reference, but they could be *consequentialist obligations and interdictions* as in the case of utilitarianism (Elster, 1990: 864). All the accounts presented in the exposition, instead, do not distinguish between social and moral norms, and see social norms as outcome-oriented. Moral norms would differ from *legal* norms too, as in the latter the self-interested motive of avoiding the punishment would be prevalent. Finally, norms cannot be convention equilibria either, since the former are sustained by a mere formal character as that of disapproval, which would not lead to any substantive consequence if transgressed, while the opposite would come about for conventions. Needless to say, a different path has been taken in this paper, as most of the works stress how conventions can be seen as a particular kind of norm, endowed with moral content.

[9] On this point see also Binmore (1993, p. 96), who argues that the matrix of payoffs of a game should not limit to the self-interested motive but should include all the relevant prompts to actions of an agent.

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

[11] One of the opponents of such an attitude-based derivation is Buchanan (1975: 132-33). He argues that such a strategy overlooks that both the component of discovering violators and punishing them involve positive costs. But, as already stressed in the exposition, the enforcement would actually be costless if the resentment hypothesis and the concept of normative expectation hold.

[12] This comes very close to Hume's account of morality:

"Conventions turn out to be a general sense of common interest; which sense all the members of the society express to another, and which induces them to regulate their conduct by certain rules. I observe that it will be in my interest to leave another in the possession of his goods, provided he will act in the same manner with regard to me. When this common sense of interest is mutually expressed and is known to both, it produces a suitable resolution and behaviour. And this may properly be called a convention or agreement

betwixt us, though with the interposition of a promise; since the actions of each of us have a reference to those of the other, and are performed upon the supposition that something is to be performed on the other part" (Hume, 1740: III.ii.2).

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

[14] To be sure, that presented here is only one of the possible approaches to the subject in a growing field of literature. Another, more refined, account is provided by Robert Sugden himself (1998b), who adds to this context a situation of uncertainty over the type of the opposing player. For a comprehensive approach, see Young (1993 and 1998), who puts forward a refinement of the concept of ESS, which must also be robust to random shocks. On questions regarding the stability of equilibria in an evolutionary context, see Skyrms (1997), who argues that in games with more than three strategies available to players, the usual mechanism of the replicatore dynamics fails to converge to steady states but generates chaotic trajectories and strange attractors.

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