

ISSN 1397-4831

WORKING PAPER 03-10

Anders Poulsen and Gert Tinggaard Svendsen

Rise and Decline of Social Capital
- Excess Co-operation in the One-Shot Prisoner's
Dilemma Game

Department of Economics
Aarhus School of Business

Rise and Decline of Social Capital

Excess Co-operation in the One-Shot Prisoner's Dilemma Game

Anders Poulsen and Gert Tinggaard Svendsen¹

Abstract:

In order to explain excess co-operation in the one-shot prisoner's dilemma game, we first question the standard assumption of stable and selfish preferences by introducing the concept of social capital. This analysis leads to a model that explains excess co-operation through an accumulation of social capital. We allow preferences and norms to change over time and hence endogenise them. Our results depend on individuals being able to acquire enough information to allow them to ascertain the social norm that fellow individuals subscribe to. However, our results do not depend on any assumptions about group size and no ostracism is required to generate co-operation. Our model gives three hypotheses about the emergence and stability of social capital and co-operation in society. First, neither unconditional cooperation nor universal defection is stable; second, for co-operation to be stable some individuals must be reciprocal in nature; third, we identify a 'rise and decline of social capital' which gives a cyclical pattern of co-operation in society. These findings may form the basis for future theoretical and empirical research.

JEL classification: A12, C71, D23, D60, D70, Z13.

Keywords: Social capital, co-operation, selfishness, reciprocity, altruism, preference dynamic, Prisoner's Dilemma.

1: INTRODUCTION

¹ Contact address: Department of Economics, the Aarhus School of Business, Prismet, Silkeborgvej 2, DK-8000 Aarhus C, Denmark. E-mails: gts@asb.dk, and aup@asb.dk. We acknowledge financial assistance from the Danish Social Science Research Foundation and helpful comments from our colleagues. Special thanks to Martin Paldam, Christian Bjørnskov and the late Mancur Olson.

Why does excess co-operation occur in the one-shot prisoner's dilemma game? Economists have long wondered why people tend to co-operate more than they should according to standard economic theory. In several experiments, one-half of the first-mover in anonymous sequential Prisoner's dilemma games chose to trust their partners. Moreover, three-quarters of second-movers reciprocated this trust, thus co-operating rather than defecting, even though the latter would have maximized their money earnings. Thus, experimental and everyday observations indicate that people tend to co-operate and follow a set of social norms, even though, in terms of short-run monetary gains, it would be rational for the individual to defect, see Van Winden (2002), McChesney (2001), Schram (2000), Schneider and Wagner (1998), Schneider et al. (1997), Ahn et al. (2001) and Sobel (2002). For example, Zak and Knack (2001) present experimental evidence on Prisoner's dilemma games which reveal a surprising amount of trusting behaviour and thereby confirm the results from Berg et al. (1995), and Smith (1997).

Our contribution is to fill the gap in the literature concerning excess co-operation by pointing to the importance of *social capital*. Social capital may be defined as the ability to co-operate in a group thus bringing about voluntary and informal collective good provisions (Coleman, 1988; Paldam, 2000; Svendsen, 2003). The existing literature on social capital has, to our knowledge, not dealt with the issue of how social capital accumulate thus determining excess co-operation. Rather, it has focused on measurement techniques and data collection, see e.g. Paldam and Svendsen (2002).

We deal with the above-mentioned gap in the literature by developing a model giving us three main hypotheses about how social capital and co-operation can be generated in a single large group (indeed the entire society), *even* when there is no repeated interaction and no possibility of explicit punishment (say ostracism). Overall they suggest that a society will, at any point in time, find itself in one of several possible situations, or phases. These are: Universal and robust co-operation, universal but increasingly fragile co-operation, increasing defection and, finally, increasing co-operation. Indeed, we will observe a cyclical pattern, in the sense that society will move through each of these phases, and then start over again. The length of time that society will spend in the 'good' situation, where co-operation is universal, will depend on the extent to which society, through its institutions and education of its young, can maintain a social norm of reciprocity and prevent one of altruism from spreading.

Research on social capital has been classified into four groups: The anthropological, the sociological, the economic and the political science approach.² Our approach, is interdisciplinary and combines, we believe, the strongest features of some of these groups. In our view individuals can be fruitfully regarded and modeled as rational, or utility-maximizing, as in the economic approach, but their preferences should not be viewed as stable, or given, and these preferences are not necessarily selfish. People may be reciprocal or altruistic (or even spiteful and competitive).³ We believe that this approach is able to deal both with the criticism that economics makes man too "under-socialized" and the criticism that sociological approaches make him "over-socialized" (see the discussion in Coleman (1988)).

Formally, our approach is related to models that analyze how individuals' preferences may change over time in the Prisoner's Dilemma game.⁴ Ockenfels (1993) shows that cooperation is evolutionarily stable as long as individuals are sufficiently likely to recognize other individuals' preferences; this recognition allows the Reciprocators to 'reward' other Reciprocators and to 'punish' the Materialists. The same basic intuition is present in our model. However, Ockenfels studies fewer preferences than we do. Precisely, he does not allow individuals to be 'altruistic'. This is a crucial difference. In Guttman (2000) two preference types compete against each other: A Reciprocator type and a Selfish type. However, as in Ockenfels, there is no altruistic type in his analysis. In our approach, on the other hand, the possibility of altruism is very important for the robustness of co-operation.

Fershtman and Weiss (1998) explore the role of social status in bringing about cooperation in the PD game. In their model, individuals get status from being more cooperative than the average and individuals care to a differing extent about status relative to money payoffs. They show that for cooperation to establish itself, the concern for social status must be sufficiently strong (such that socially minded people can co-ordinate on the cooperative outcome), but not too strong (then socially minded individuals act like altruists who can be exploited by asocial people). However, it is, as above, relevant to point out that Fershtman and Weiss only allow for two preference types. Indeed, in their model the set of feasible types is either a 'Selfish' or a 'Reciprocal' type, or a 'Selfish' and 'Altruist' type.

² See, for example, OECD (2001).

³ See Fehr and Fischbacher (2002), for a taxonomy of 'social preferences'.

⁴ Early contributions are Witt (1986) and Güth and Yaari (1992).

The rest of the article is organized as follows. In Section 2 we first question the standard assumption of stable and selfish preferences. Section 3 then introduces the concept of social capital. Section 4 sets up a model that analyses the interaction between social norms, social capital and the emergence of co-operation. All the formal results are given in the Appendix.

2: THE STANDARD VIEW OF PREFERENCES

The standard view in economics is that preferences are *stable* (or given) and *selfish*. Individuals are concerned only with their own monetary payoff (see, for example, Becker and Stigler (1976)). Kelman (1987, p. 81) criticises the traditional assumption of selfishness in economics by claiming that the role of a ‘public spirit’ also plays an important role in motivating individual behaviour. Economic analysis cannot just be applied to politics, because people do not necessarily act everywhere as they do in the market place.⁵ People are not always out only for themselves in money-seeking ways like ‘pigs at the trough’ – an understanding which leads to the following views:

Do you want to understand why government officials behave the ways they do? All you need to know is that they are trying to maximize the budgets of their agencies. Do you want to understand what drives politicians? All you need to know is that they want to be re-elected. Do you want to understand legislation? Just see it as a sale of the coercive power of government to the highest bidder, like a cattle auction (ibid., p. 81).

Kelman says that this account of the political process is a caricature of reality because it ‘...ignores the ability of ideas to defeat interests and the role that public spirit plays in motivating participants...’ and that ‘...a norm of public spiritedness in political action – a view that people should not simply be selfish in their political behaviour – is crucial. The public choice school is part of the assault on that norm’ (ibid.). Crucial to any ability to maintain public spirit is the continuing existence of a social norm that declares it appropriate for people to try to do the right thing in public behaviour, and inappropriate for them simply to seek to advance their personal interests. If the norm of public spirit dies, our society would look bleaker and our lives as individuals would be impoverished. The Robinson Crusoe idea that people are best conceived of as isolated units because is inadequate

⁵ And, in the market place motivations like reciprocity and fairness also play an important role. See, for example Kahneman, Knetsch and Thaler (1986) and Fehr, Kirchsteiger and Riedl (1993).

psychology as well as unfashionable economics. That is the tragedy of 'public choice' (ibid., pp. 93–94).

In the same vein, Cram (1998, p. 77) notes that in everyday life, 'people don't stop at every choice they make in their life and think about what will maximize my self-interest' Rather, they mostly follow socially defined rules which may not necessarily be in their own short-run material self-interest. These views are overall consistent with the political science literature that depicts preferences as endogenous, i.e., that preferences develop within the institutional setting. In contrast, the overall economic literature perceives preferences as exogenous, i.e., given and independent of institutional set-ups. This point is the main political science objection against economics, namely that preferences are non-stable and not necessarily selfish (Green and Shapiro, 1994).

We will seek to address these limitations of the standard view on preferences by introducing the concept of social norms and social capital in the following section.

3: SOCIAL CAPITAL

People differ in the perceptions, motives and social norms that they 'carry' with them in society. The social norm that a person adheres to is expressed through the person's willingness to cooperate or defect in games like the Prisoner's Dilemma game. We will say that a *social norm creates social capital when it allows the individuals carrying the social norm to rationally co-operate in the one-shot Prisoner's Dilemma*. This is in line with the definition of social capital, given in e.g. Paldam (2000).

In the presence of social capital and trust, fewer will commit crime, free-ride or ignore the rules in a contract. This is in stark contrast to the traditional political economy approach, which tends to ignore social norms and social capital when dealing with human nature. Here, if the risk of detection (and the cost associated with punishment) is lower than the benefits from committing a crime, a person will commit the crime. Anyone will rob an old lady in the street if it pays. In the absence of police and court systems, anarchy will occur (Svendsen, 2003).

Hobbes is the first to offer the classical solution to discipline human nature, which is that of third-party enforcement and the establishment of strong formal institutions. Because the

weakest member in society is capable of killing the strongest, a social contract is needed. This contract is enforced by the totalitarian king; 'Leviathan', who shall protect people against each other. If there were no common power to restrain individuals, no law and no law-enforcement, every man would constantly be open to violent invasion of his life and property. Anarchy means '...continually feare, and danger of violent death; And the life of man, solitary, poore, nasty, brutish, and short', Hobbes ([1651] 1985, p. 186). This is in stark contrast to the view of Rousseau, who describes life in nature as harmonic and peaceful. The problem is that people have been removed from the state of nature. People are never corrupted and have goodwill, but they are often deceived (Rousseau, [1762] 1993, p. 203) (cited from Paldam and Svendsen, 2000).

The key element in social capital is, in contrast to this traditional political economy approach, that society, via voluntary informal institutions and social norms, can enable individuals to *trust* that other individuals will act co-operatively. This will happen *even* in the absence of a 'Leviathan'. Such trust assures an individual that he will not be taken advantage of by another individual, even if the latter might get an economic net benefit from doing it. Even if it pays economically to commit a crime, free-ride or ignore the rules in a contract, fewer will do it in the presence of trust because social norms tell them not to do so. Another key idea in the social capital approach is that the community members' preferences can be affected and shaped, due to social norms and social pressures (see Becker, 1996 and Green and Shapiro, 1994, for further discussions on unstable preferences). Overall, formal mechanisms are not necessary, since the social capital in the community does the job.

When social capital in this way reduces the need for a Leviathan, i.e. third-party monitoring and enforcement of contracts, more transactions can take place at a lower cost. In this way, trust 'lubricates' society. For example, it may be argued that a firm lowers transaction costs by having numerous informal transactions taking place that are not formally sanctioned (see Coase, 1937). In this way, a society with members that trust each other may be capable of accomplishing more economic growth than a similar society without trust. Consequently, social capital could be a new production factor, which must be added to the conventional concepts of human and physical capital. In relation to macroeconomic policies, it follows that if people hold a high level of social capital, they will trust each other and elected decision-makers as well. By that, the build-up and maintenance of social capital will ease policy-making and make it more effective since less monitoring is necessary when a high

level of social capital is present. This point is illustrated by a new survey on shadow economy activities by Schneider and Enste (2002). Here, one could argue that less monitoring of shadow economy activity and tax evasion was needed in countries with high levels of social capital because people generally cheat less.

Given all these observations, our key aim in the following section is to explore the question: To what extent will a social norm that allows individuals who have internalized this norm in their preferences to rationally co-operate, even in a one-shot situation and in a large community without neither formal nor informal enforcement (such as ostracism)?

4: A SIMPLE MODEL OF EMERGENCE AND VIABILITY OF SOCIAL CAPITAL

In this section we study a simple formal model, in order to shed some light on how social capital is created and the extent to which it can be maintained.

4.1. The basic set-up

We imagine a large city, where people occasionally meet each other randomly on the central square. They meet in pairs, and then play the *one-shot* PD game, where they simultaneously choose between Cooperate (*C*) and Defect (*D*). The matrix below gives the money payoffs for the Prisoner's Dilemma (PD) game.

	C	D
C	<i>1,1</i>	<i>b,a</i>
D	<i>a,b</i>	<i>0,0</i>

Figure 1: The Prisoner's Dilemma game (the number in each cell are money payoffs).

We have $a > 1$ and $b < 0$. If individuals are only concerned about maximizing their own money payoffs, as in the standard economic model, then the outcome will be that each individual chooses to defect (*D*). Each individual then earns zero money. However, could the individuals instead have established mutual cooperation (each individual chooses *C*), each individual would have earned a money payoff of unity.

In our model, social norms, embedded in individuals' preferences, may induce them to behave in a way that differs from the one normally assumed. We will consider the following three social norms:

1) The *Reciprocal social norm (R)*: Reciprocate what the other opponent does, or is expected to do: If the opponent is going to, or is thought to play *C*, then an individual adhering to the Reciprocal norm will play *C*, too. Similarly, if the opponent is going to, or is suspected to play *D*, a Reciprocal individual will play *D*, too. Our interpretation of such behavior is that the individual has internalized a norm of *reciprocity*: If your opponent is (or is thought to be) kind, then be kind, too; and if the opponent is (thought to be) unkind, be unkind, too. We note that a Reciprocal person does *not* always maximize her monetary earnings: If the opponent plays *C*, she plays *C*, too, whereas a choice of *D* would have given her more.

Reciprocity has ancient roots. The following quote [from Fehr and Gächter (2000)] is from the *Edda*, a 13th century collection of Norse epic verses:

"A man ought to be a friend to his friend and repay gift with gift. People should meet smiles with smiles and lies with treachery."

Reciprocity is a vigilant and attentive social norm: Behavior is *conditioned* on the opponent's identity.⁶

2) The *Selfish social norm (S)*: A person adhering to this norm *always* plays Defect. Our interpretation is that she obeys an (a)social norm of always seeking to maximize her monetary return. She is the 'standard' type of individual assumed in most of economics and game theory. In the style of the *Edda*, the Selfish type would subscribe to the following maxim:

"Don't rely on anyone! Receive, but never give gifts. Never smile and lie whenever it helps you."

⁶ For good surveys of reciprocity and other 'social preferences', we refer the reader to Fehr and Gächter (1998) and Fehr and Fischbacher (2002).

3) The *Altruist social norm (A)*: An individual following this norm *always* chooses to cooperate. An interpretation is a norm of altruism: Always seek to maximize total (or just the opponent's) money surplus. The Altruist type could be interpreted as a certain variant of Kant's Categorical Imperative: Treat other's as you want them to treat you. The Altruistic type is a form of *unconditional altruism*. Inspired by the verse from the *Edda* quoted above, we could write: "Be good to everybody! A man ought to be a friend to everybody, and to give gifts. He should smile and never lie."

The Reciprocal type *distinguishes* between opponents, on the basis of their preferences (i.e., social norms): If the opponent is thought to be 'kind' (play *C*), she plays *C*, too; otherwise she acts in an unkind manner (plays *D*). We may say that the Reciprocal type is *conditionally kind*. The Altruist type, on the other hand, is *unconditionally kind*: She plays *C* against all opponents. This is, as we shall see below, a crucial distinction, and it may affect the robustness of cooperation in a, perhaps at first sight, unintuitive way.

We will assume that when two individuals meet, each individual learns what the opponent's social norm, or preferences, is. Strictly speaking, of course, this is impossible: Individuals cannot 'see' inside other persons' heads. However, several authors, such as Robert Frank, 1988, have argued that it is often possible to correctly deduce people's characteristics, and underlying motivations, from physical tell-tale signs, such as facial expressions and body posture. In our model people could then infer from such physical signs whether a person intends to unambiguously cooperate (an *A* type), defect (type *S*) or is ready to reciprocate cooperation and defection (type *R*). Another possibility is that people can ask around in the community about an opponent's previous behavior in past encounters and from that form an opinion about what social norm the individual subscribes to. Yet another possibility is that individuals base their evaluations of other individuals' preferences using indicators such as income, skin color, area of residence, and so on, *and* that these are correlated with the social norms followed. We leave for future research the task of incorporating such, and other, realistic features into models of social norms and social capital.⁷ Whatever the specific mechanism, we assume individuals on average correctly deduce their opponents' internalized social norms.

⁷ For a formal model, see Kandori (1992).

We can now consider what happens when individuals with different social norms or preferences meet and play the one-shot PD game given above. Let us focus on an encounter between two *R*-types. Here there are two possible outcomes: One where both are 'kind', i.e., choose *C*, and another where both are 'unkind', i.e., choose *D*. Precisely, both the (*C,C*) outcome and the (*D,D*) outcome are Nash equilibria.⁸ We will assume that two Reciprocators can establish the cooperative (*C,C*) outcome. Our interpretation is that each *R*-individual *trusts* that the *R*-opponent will play *C*, and hence the individual plays *C*. How does this trust arise? When two *R*-types meet, they see that they subscribe to the same social norm, one of reciprocity. This is not in itself enough to get co-operation, since (*D,D*) is also a Nash equilibrium. However, the (*C,C*) outcome is better in monetary terms for *both* individuals than the (*D,D*) outcome. We assume that this persuades each individual that the opponent will indeed co-operate. It follows that each *R*-type earns a monetary payoff of unity.

The money payoffs arising from other encounters are quite straightforward to derive.⁹ We thus obtain the matrix below, which contains the money payoff earned by a type in encounters with the other types. Cell (*i,j*) gives the money payoff earned by type *i* when meeting an opponent of type *j*, where $i,j=A, R, S$.

⁸ To see that (*D,D*) is a Nash equilibrium, note that an *R*-individual prefers *D* if the opponent plays *D*. It follows that if two *R*-types play *D*, each individual has no incentive to deviate. The same argument shows that (*C,C*) is a Nash equilibrium. There is, in addition to these two equilibria, also a mixed Nash equilibrium, where each *R*-type mixes between *C* and *D*, but we will ignore this possibility here.

⁹ *A-A*: Here both persons cooperate. Thus each gets a money payoff of unity. *A-R*: The *A*-type plays *C*; the *R*-type does the same, since the *R*-type observes that the opponent will play *C* and hence her best reply is to do the same.

A-S: Here each individual chooses *D*. The reason is that the *S*-type always plays *D*, and given this the *R*-type's best reply is to play *D*, too. *R-S*: Since the *S*-type always plays *D*, the *R*-type does the *same*. It is here we see the flexibility of a norm of reciprocity: Sometimes it instructs an individual to play *C* (when meeting an *R*- or an *A*-type), other times the result is *D* (when meeting an *S*-type). *S-S*: Here both individuals defect.

	<i>A</i>	<i>R</i>	<i>S</i>
<i>A</i>	<i>l</i>	<i>l</i>	<i>b</i>
<i>R</i>	<i>l</i>	<i>l</i>	<i>0</i>
<i>S</i>	<i>a</i>	<i>0</i>	<i>0</i>

Figure 2: The money payoffs realized in meetings between the three types of individuals.

A = Altruist; *R* = Reciprocal; *S* = Selfish.

We see that an altruist person performs well against other altruists and well against reciprocal individuals (they all co-operate), but is exploited by a Selfish person. The Reciprocator performs well against other reciprocal and altruist individuals, like the altruist, but is not exploited to the same extent as the altruist when meeting a selfish individual. Finally, the selfish person earns a lot of money when meeting an altruist and otherwise mediocre.

We wish to endogenously determine how many people will internalize what kinds of social norms in their preferences. To do so, we postulate a dynamic process through which individuals' preferences may change over time. Our key assumption will be that those social norms or preferences who give their 'users' higher-than-average *money* payoffs tend to be internalized by more individuals over time.¹⁰ An interpretation is that, in order to survive in the economic system, a social norm needs to give its followers a sufficiently high level of material welfare. This assumption, that 'only money matters', does not however a priori bias the analysis towards the survival of materialistic preferences: *Any* sort of social norm or preference that leads individuals to a materially superior, or just reasonable, behavior will prosper. Thus, if individuals with, say, Reciprocal, preferences earn more money than those with materialistic preferences, the population frequency of the former will tend to increase at the expense of the latter.

Formally, let x_i , where $i=A, R, S$, indicate the population proportion of individuals of type i . That is, we have $0 \leq x_i \leq 1$ and $\sum_i x_i = 1$. Let $x = (x_A, x_R, x_S)$ denote the population distribution of

¹⁰ This methodology is called the 'indirect evolutionary approach' (see e.g. Güth and Yaari (1992)).

the three preference types. Moreover, let $\pi(i,x)$ denote the expected payoff to an i -type at the population x .¹¹ Finally, we denote by $\pi(x,x)$ the average payoff in the population.

The growth rate of the proportion of individuals of type i is then given as

$$\dot{x}_i / x_i = \pi(i,x) - \pi(x,x).$$

This is the so-called Replicator Dynamic (see e.g. Weibull (1995)). It says that the growth rate of individuals with preference $i=A,R,S$ grows if these individuals earn above-average *money* payoff. We wish to describe the dynamic of preference evolution and to find those population states that are *stable* for this dynamic. Stability means, roughly, that the population cannot be invaded by individuals following other social norms.¹²

4.2. Analysis: Three hypotheses about social capital and co-operation

Let us first consider the population where all individuals are Selfish, i.e., they always play Defect. Is this population stable – that is, is it possible for some individuals to 'mutate' to another preference, i.e., subscribe to another social norm, that will give them more money than the incumbents? Suppose indeed a small group of individuals starts to evolve reciprocal preferences. Let us compare the performance of a Selfish incumbent and the 'new' Reciprocal individual. Each meets essentially only Selfish individuals and the Selfish person gets zero in such an encounter, as does the Reciprocal mutant. Thus the mutant performs exactly as well as the incumbent, so nothing prevents more mutants from slowly entering the population. However, the Reciprocal mutant performs strictly better against itself than a Selfish individual. This implies that as soon sufficiently many Reciprocal mutants have entered the population, the mutants will earn strictly more money than the Selfish incumbents. The reciprocal social norm is then more successful in money terms than the selfish one, so the proportion of Reciprocal individuals increase, at the expense of the Selfish individuals and the dynamic takes us away from the all-Selfish population.

Consider next the population where all individuals pursue the norm of cooperating unconditionally with everybody, i.e., all individuals are Altruists. Can this social norm persist? Unfortunately, the answer is no, since a payer who deviates from the existing social

¹¹ This expected payoff to, say, the R -type at x is $\pi(R,x)=x_G(1) + x_R(1) + x_M(0)$.

norms and evolves a Selfish preference earns more money than the incumbents and hence will be followed. Thus even if we so lucky as to start there, we should not expect to see a purely altruistic society for a very long time.

Hypothesis 1: (a). A society where all individuals have internalized the Selfish social norm will not be observed in the longer run. (b). The same will be true for a society where all individuals follow the Altruist norm.

The formal basis for the hypothesis can be found in the Appendix. A society of selfish individuals, where all people defect in each other, will not be resistant to entry by reciprocally individuals. And an altruistic social norm of cooperating *unconditionally* with everybody cannot be sustained in a social or economic system where money matters for performance.

Let us now consider potentially 'mixed' populations, consisting of Reciprocal and Altruist types. In such a population everybody cooperates when they meet, no matter their type (*A* or *R*). The potential invaders of this cooperative state of affairs are, not surprisingly, a group of individuals subscribing to the *S*-type. However, as long as most of the incumbents are *R*-types, these *S*-mutant cannot invade: Each mutant gets mostly zero money payoff (since it mostly meets the *R*-individuals), while the incumbents get a money payoff of unity from their co-operation.

However, notice that the *A* and *R*-type perform equally well in the population (there is co-operation in all encounters, so they earn the same money payoff). This, in turn, gives rise to the phenomenon of 'evolutionary drift' (see e.g. Binmore and Samuelson (1994)): The proportions of *A* and *R* types may slowly and unnoticeably change, due to stochastic background factors that we have not explicitly modeled. Moreover, the higher the proportion of *A*-types in the population, the higher the money payoff to any *S*-mutant (recall that the mutant will get the high *a* money payoff whenever meeting an *A*-type). If, due to such evolutionary drift, the proportion of *A*-types exceed a certain limit, an *S*-mutant *can* invade the population. Then the pattern of universal cooperation is broken and defection emerges.

¹² For more precise definitions, we refer the reader to e.g. Weibull (1995).

These observations are collected in the proposition below:

Hypothesis 2: A society where some individuals are Reciprocal and the rest are Altruists will be stable if there are not too many Altruist individuals in the population. There will be a critical threshold proportion of Altruist individuals, such that the population is stable, as long this threshold is not exceeded. In our model, this threshold is given by $1/a$. If the threshold is exceeded, some individuals will adopt a Selfish norm and defection will increase.

The formal proof can be found in the Appendix. The reason for the evolutionary drift mentioned above is essentially that Reciprocators co-operate with the Altruists; the Reciprocators do not 'punish' the Altruists for being different.

Our two hypotheses are illustrated in the figure below (where we have set $a=2$ and $b=-1$):

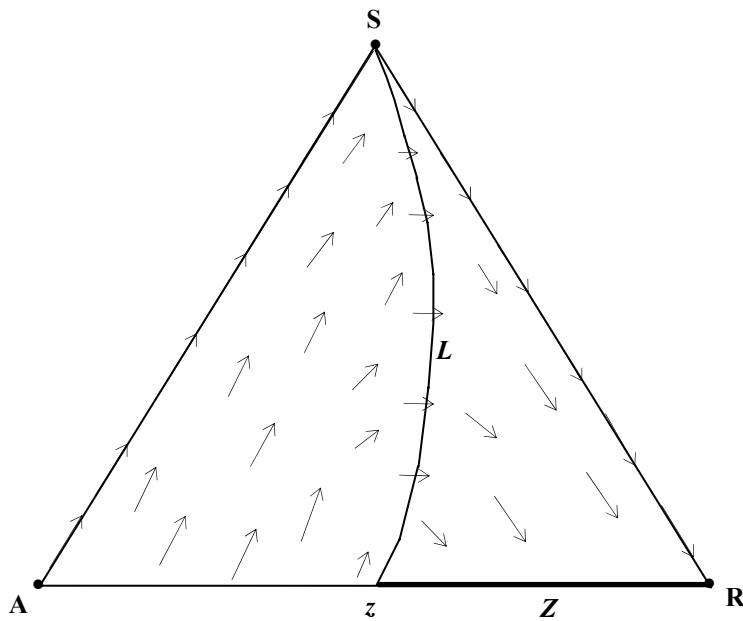


Figure 3: The evolution of social norms and social capital: Illustration of Propositions 1, 2 and 3. ($a=2$ and $b=-1$).

The triangle contains all the possible populations. The corner labeled " i " is the population where all individuals are of type i , where $i=A,R,S$. Along an edge $i-j$ ($i,j=A,R,S$) individuals are either of type i or type j (individuals of the third type are 'extinct'). Inside the triangle all

three types of individuals are present in the population. The arrows indicate the direction of change of the distribution of social norms in society.

Consider first the S - R edge. Here the dynamic takes the population away from the all- S population. This is what gives part (a) of *Hypothesis 1*. Similarly, the population is taken away from the all- A population; this is part (b) of *Hypothesis 1*.

The line segment Z , along the A - R edge, gives the populations consisting of R -types and A -types that will be stable (cf. Hypothesis 2). If the proportion of A -types exceeds the critical value (point z in the figure), then S -mutants can enter the population and the population moves out into 'open sea', where all three preference types are present. This gives us *Hypothesis 2*. Note it is theoretically possible that the population can continue along the A - R edge to the right of point z , as long as no individuals 'mutate' to the S norm. However, sooner or later, such a mutation will take place and then the population leaves the A - R edge and is taken inside the triangle.

From this point onward the population proportion of S -types will increase and the proportion of A -types and R -types will drop. However, sooner or later, there will be so few A -types in the population that the money earnings of the S -types starts fall below the average; then the proportion of R -types start to increase and that of S -types start to fall (the curve L , which connects the S corner with the point z , is crossed in a leftward movement). Then the population moves towards the line segment Z again. And, once there, *the story can repeat itself*.

In a population composed initially only of R -types (the population is at the R corner), how long time will it take for altruism to spread such that the population moves leftward along line segment Z and reaches the critical threshold level at the point z ? Due to the fact that, as already mentioned, the A and R -individuals perform exactly well, the population may move back and forth along line segment Z . Given that both leftward and rightward movements are possible, the population will *eventually* reach the point z , after which any leftward movement will lead the population away from segment Z and into the interior of the triangle. However, the factors that determine how *fast* this will take place are not explicitly modeled in our model. However, it is not difficult to think of what is important: The general institutional set-up in society; culture, the values that children obtain from their teachers and parents etc. In general, the ability of society to sustain a reciprocal social norm among its members and to inflict, say, a moral cost, on adopting a norm of unconditional kindness.

We see that there is a *cyclical pattern of building up, erosion, and destruction of social capital*: An initial dominance by the Reciprocal social norm of cautious and conditional co-operation gives maximum social capital (the *R* corner); slowly, however, a social norm of unconditional co-operation (the Altruist norm) can emerge, but social capital and cooperation is still at its maximum since everybody co-operates (we move leftward along line segment *Z*); eventually, however, the critical threshold of altruism is exceeded (cf. Hypothesis 2), and this leads to an invasion by individuals adhering to the Selfish social norm; at this point an erosion of social capital starts, with the *S* norm increasing and the *A* and *R* norms diminishing in popularity (we move inside the triangle, toward corner *S*); we observe an increasing erosion of social capital. However, at some point there will be too many Selfish individuals and too few Altruist individuals that the Selfish individuals start to perform worse than the other type (in Figure 3 the curve *L* is crossed). We will see a rehabilitation of the *R* norm, a fading of the *S* norm, and a strengthening of the *A* norm; in other words, social capital is re-created. Eventually all Selfish individuals will disappear and we have, once more, maximum social capital and universal co-operation (the population 'lands' on line segment *Z*). From this point and onwards, the whole dynamic process may repeat itself.

In summary, society can go through three qualitatively different phases: "Cooperation, but risk of Increasing Decadence: Potential Destruction of Social Capital " (moving leftwards along line segment *Z*); "Spread of Selfishness and Destruction of Social Capital" (moving from point *z*, or from some other point on the edge to the left of point *z*, into the interior of the triangle towards point *M*); "Cleansening/Recovery: (Re)Creation of Social Capital" (crossing the *L* curve and approaching line segment *Z*).

We summarize all these observations as the following hypotheses:

Hypothesis 3: (a). A society where all, or most people, are Selfish will over time experience increasing degree of co-operation. (b). A society that has established universal co-operation can not expect to stay there indefinitely. The time during which it will sustain universal cooperation, however, depends on the extent to which society can put a stop to the spreading of altruism. (c). A society that has been invaded by selfishness will go through a period of increasing defection, until cooperation again establishes itself.

The figure below illustrates the dynamic pattern of defection and co-operation. In the figure, the degree of cooperation in society, c , is measured along the vertical axis and time is measured along the horizontal axis.¹³

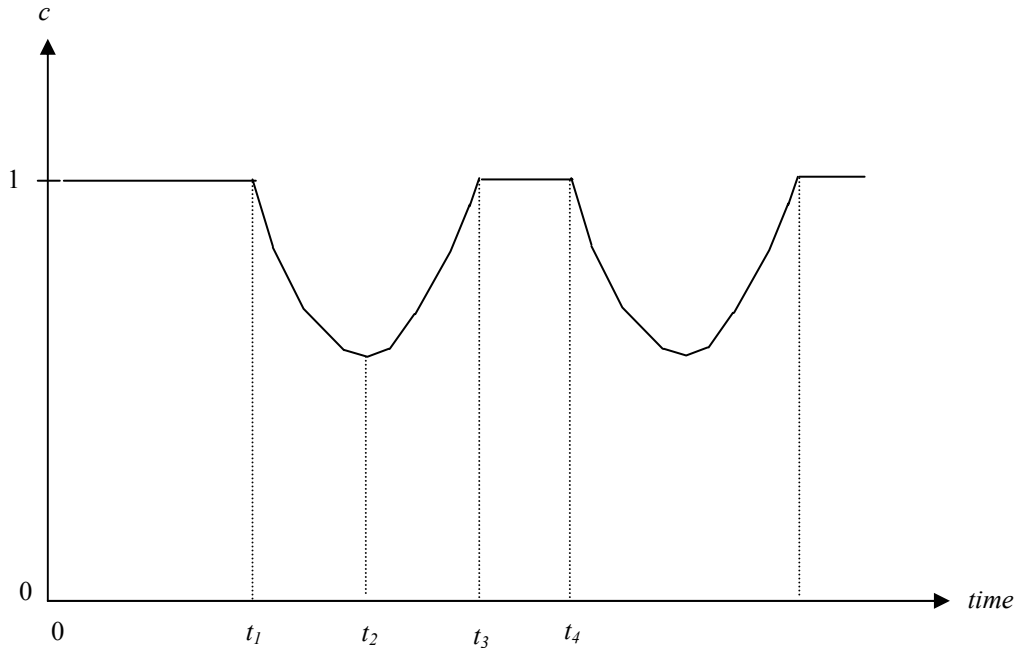


Figure 4: A cyclical evolution of co-operation and defection.

We start out in the situation where all individuals are reciprocal (corner R in Figure 3); thus there is universal co-operation: $c=1$. Over time, individuals start to subscribe to the Altruist norm, i.e., we move leftwards along line segment Z in Figure 3). After some time (time t_1), the critical share of Altruists is reached. Whenever this is exceeded, any mutation of individuals to the S norm will be materially successful and be followed by more persons. Thus the proportion of selfish individuals increases, and hence the frequency of defection, increases, too. This continues until there are so few Altruists that the Reciprocators start to perform better than the Selfish individuals (time t_2 , corresponding to the crossing of the L curve in Figure 3); then the frequency of co-operation starts to increase again until it is fully re-established at time t_3 (which corresponds to arrival on line segment Z). Society may then again spend some time here until (time t_4) there are again too many altruists, allowing the selfish norm to become present. The pattern then repeats itself. We see that the time society

¹³ The degree of co-operation is simply measured as the relative frequency of (C,C) outcomes.

enjoys universal co-operation may vary (in the figure distance $0-t_1$ exceeds distance t_3-t_4), due to random noise and due to society's potentially varying ability to fend off altruism. The more able society is to prevent the spreading of altruism, the longer the periods of universal co-operation.

4.3. Summary:

We can summarize the hypotheses of our model as follows: Social capital and co-operation will establish itself in a 'hostile' world of materialism and selfishness whenever individuals following a reciprocal social norm can, on average, correctly distinguish between their own type and 'aliens'. However, there will be a risk that unconditional niceness, or altruism, will erode reciprocity and eventually allow selfishness to spread. From that point onwards, society will go through a period of cleansing, where altruism and reciprocity will diminish; society will become largely materialistic. From that point and onward, however, reciprocity can establish itself.

5: CONCLUSION

It is widely acknowledged that the social capital concept is important, yet there is a need for further theoretical clarification and empirical research. We have tried to contribute to a new understanding of how it is that social capital rise and promote a society's ability to establish co-operation in a world where purely material considerations would lead to universal and inferior defection and where formal sanctions and ostracism are not available.

In order to explain excess co-operation in the one-shot prisoner's dilemma game, we first questioned the standard assumption of stable and selfish preferences by introducing the concept of social capital. This analysis led to a model that explained excess co-operation through an accumulation of social capital. We allowed preferences and norms to change over time. Our results depended on individuals being able to acquire enough information to allow them to ascertain the social norm that fellow individuals subscribed to. However, our results did not depend on any assumptions about group size and no ostracism was required to generate co-operation.

Our model generated three hypotheses identifying a rise and decline of social capital and co-operation in society. First, neither unconditional cooperation nor universal defection are stable; second, for co-operation to be stable some individuals must be reciprocal in nature; third, and partly resulting from these two hypotheses, society will go through a cyclical pattern, where social capital is first rising, then consolidating itself, then gradually weakened, and then invaded by selfishness, after which a 'cleansening' will take place that will allow social capital to rise again. Most importantly, society's ability to promote long-run co-operation depends on its ability to prevent altruism from spreading in the population. We hope that some of our findings may form the basis for future theoretical and empirical research.

6: APPENDIX

Here we provide the formal basis for our three hypotheses. The results below follow from applications of well-known results from evolutionary game theory. We refer the reader to one of the many good books on this topic, such as Weibull (1995).

Hypothesis 1: Part (a): Consider the all- S population and any arbitrarily small perturbation injecting a small proportion of R -individuals, such that the post-perturbation population contains S and R -individuals. Against this population the R -strategy performs strictly better than the S -strategy, implying that the all- S population is unstable. Part (b): This follows from the fact that the A strategy is not a Nash equilibrium for the game in Figure 2, hence unstable for our dynamic system. ■

Hypothesis 2: Let x denote any population where a proportion x_A are A -individuals and the remaining proportion, $1-x_A$, are R -individuals. At this population everybody co-operates, so average payoff is $\pi(x,x)=1$. Facing x an S -individual earns expected payoff $\pi(S,x)=x_A a$. It follows that when $\pi(S,x) \leq \pi(x,x)$, or $x_A \leq 1/a$, then x is a Nash equilibrium. Consider any x satisfying $x_A < 1/a$. To show that x is stable for the Replicator Dynamic, it suffices (see e.g. Weibull (1995)) to show that x is a so-called Neutrally Stable Strategy (NSS). This means that x satisfies the following two conditions: (i) $\pi(x',x) \leq \pi(x,x)$ for all x' and (ii). For any $x' \neq x$ satisfying $\pi(x',x) = \pi(x,x)$ we have $\pi(x,x') \geq \pi(x',x')$. We already know that condition (i) holds for all x' whenever $x_A < 1/a$. Moreover, from the same inequality we see that $\pi(x',x) = \pi(x,x)$ only for those $x' \neq x$ that assigns positive probability to strategy A and/or to strategy R . But for

any such x' we have $\pi(x,x')=\pi(x,x)=1$. Thus condition (ii) is also satisfied, so we may conclude that any x with $x_A < 1/a$ is an NSS and hence stable for the Replicator Dynamic. ■

Hypothesis 3: Here we provide the derivation of the dynamic in Figure 3. From the Replicator Dynamic the equations giving the expected payoff to the preference types are $\pi(A,x) = 1-x_S + x_S b$, $\pi(R,x) = 1-x_S$ and $\pi(S,x) = (1-x_R-x_S)a$. The average expected payoff is $\pi(x,x) = 1-2x_S+bx_S-bx_Rx_S+x_S^2-bx_S^2+x_Sa-x_Rx_Sa-x_S^2a$. Our Replicator Dynamic is therefore $dx_A/dt = x_S[1+x_Rb-x_S+bx_S-a+ax_R+x_Sa]$, $dx_R/dt = x_S[1-b+x_Rb-x_S+bx_S-a+ax_R+x_Sa]$ and $dx_S/dt = a-ax_R-2ax_S-1+2x_S-bx_S+bx_Rx_S-x_S^2+bx_S^2+ax_Rx_S+ax_S^2$. Let us now set $a=2$ and $b=-1$. We then verify that $dx_A/dt=x_S(x_R-1)$, $dx_R/dt=x_Rx_S$ and $dx_S/dt=1-2x_R-x_S+x_Rx_S$. We see that dx_A/dt is non-positive for all populations x with $x_S > 0$ and $x_R < 1$. Thus the proportion of A -individuals is falling at any x in the interior in the state space. Moreover, the proportion of R -individuals is increasing at any population x in the interior of the state space. Finally, the proportion of S -individuals is increasing [decreasing] when $x_2 < [>] (1-x_S)/(2-x_S)$. Plotting the graph of the equation $x_2=(1-x_S)/(2-x_S)$ gives the L curve in Figure 3. ■

7: REFERENCES

- Ahn, T.K., E. Ostrom, D. Schmidt, R. Shupp and J. Warker (2001), ‘Cooperation in PD games, fear, greed and history of play’, *Public Choice*, **106**, 137–155.
- Becker, G. (1996), *Accounting for Tastes*, Harvard University Press, Cambridge, Massachusetts. Cambridge University Press, New York.
- Becker, G. and Stigler, G. (1976): "De Gustibus Non Est Disputandum", *American Economic Review*, **67**, 76-90.
- Berg, J., J. Dickhaut and K. McCabe (1995), ‘Trust, reciprocity and social history’, *Games and Economic Behaviour*, **10**, pp. 122–42.
- Binmore, K. and Samuelson, L. (1994): "Drift", *European Economic Review*, **38**, 859-867.
- Coleman, J.S. (1988), ‘Social Capital in the Creation of Human Capital’, *American Journal of Sociology*, **94**, 95–120.
- Cram, L. (1998), ‘The EU institutions and collective action: Constructing a European interest?’, in Greenwood and Aspinwall (eds.), *Collective Action in the European Union*, London: Routledge, pp. 63–80.

- Fehr, E. and Gächter, S. (1998): "Reciprocity and economics: The economic implications of Homo Reciprocans", *European Economic Review*, **42**, 845-859.
- Fehr, E. and Gächter, S. (2000): "Fairness and Retaliation: The Economics of Reciprocity", *Journal of Economic Perspectives*, **14**, 159-181.
- Fehr, E. and Fischbacher, U. (2002): "Why Social Preferences Matter - The Impact of Nonselish Motives on Competition, Cooperation, and Incentives", *Economic Journal*, **112**, C1-C33.
- Fehr, E., Kirchsteiger, G. and Riedl, A. (1993): "Does fairness prevent market clearing? An experimental approach", *The Quarterly Journal of Economics*, 437-459.
- Fershtman, C. and Weiss, Y. (1998): "Why do we care what others think about us?", 133-151 in Ben-Ner, A. and Putterman, L. (eds.): *Economics, Values and Organization*, Cambridge University Press.
- Frank, R. (1988): *Passions Within Reasons*, W.W. Norton & Co.
- Green, D.P. and I. Shapiro (1994), *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*, Yale University.
- Guttman, J. M. (2000): "On the evolutionary stability of preferences for reciprocity", *European Journal of Political Economy*, **16**, 31-50.
- Güth, W. and Yaari, M.: (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game". In Witt, Ulrich (ed.): *Explaining Process and Change - Approaches to Evolutionary Economics*, Ann Arbor, MI: University of Michigan Press.
- Hobbes, T. ([1651] 1985), *Leviathan*. London: Penguin Classics.
- Kahneman, D., Knetsch, J. and Thaler, R. (1986): "Fairness and the Assumptions of Economics", *Journal of Business*, **59**, S285-S300.
- Kandori, M. (1992): "Social Norms and Community Enforcement", *Review of Economic Studies*, **59**, 63-80.
- Kelman, S. (1987), 'Public Choice and Public Spirit', *Public Interest*, **87**, 80-94.
- McChesney, F.S. (2001), 'Rent seeking and rent extraction', in William F., W.F. Shughart II and L. Razzolini (eds.), *The Elgar Companion to Public Choice*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 379-95.
- Ockenfels, P. (1993): "Cooperation in prisoner's dilemma", *European Journal of Political Economy*, **9**, 567-579.
- OECD (2001): *The Well-being of Nations – The Role of Human and Social Capital*.

- Paldam, M. (2000), 'Social Capital: One or Many? Definition and Measurement', *Journal of Economic Surveys*, **14** (5), 629–53. Special Issue on Political Economy.
- Paldam, M. and G.T. Svendsen (2000), 'An essay on social capital: Looking for the fire behind the smoke', *European Journal of Political Economy*, **16**, 339–66.
- Paldam, M. and G.T. Svendsen (2002), 'Missing Social Capital and the Transition in Eastern Europe', *Journal of Institutional Innovation, Development and Transition*, **5**, 21–34.
- Rousseau, Jean-Jacques, (1762), *Du Contrat Social*, Paris: Hachette/Pluriel, 1978.
- Schram, A. (2000), 'Sorting out the Seeking: The Economics of Individual Motivations', *Public Choice*, **103** (3/4): 231–58.
- Schneider, F. and A. F. Wagner (1998), *Emission Trading and Environmental Taxes as Efficient Flexible Instruments for European Climate Policy: Remarks from an Economist's Perspective*, University of Linz.
- Schneider, M., P. Teske, M. Marschall, M., Mintrom and C. Roch (1997). 'Institutional Arrangements and the Creation of Social Capital: The Effects of Public School Choice', *American Political Science Review*, **91** (1), 82–93.
- Schneider, F. and Enste, D.H. (2003), *The Shadow Economy*. UK: Cambridge University Press.
- Sobel, J. (2002), 'Can we trust social capital?' *Journal of Economic Literature*, **XL**, 139–54.
- Smith, V.L. (1997), 'The two faces of Adam Smith', Working paper, University of Arizona.
- Svendsen, G.T. (2003): *Political Economy of the European Union: Institutions, Policy and Economic Growth*. Edward Elgar, Cheltenham, UK. In press.
- Van Winden, F. (2002), 'Experimental investigation of collective action?' Paper presented at the 2002 Annual Meeting of the European Public Choice Society in Belgrate, Italy.
- Weibull, J.W. (1995): *Evolutionary Game Theory*, MIT Press.
- Witt, U. (1986): "Evolution and Stability of Cooperation without Enforceable Contracts", *Kyklos*, **39**, 245-266.
- Zak, P.J. and S. Knack (2001), 'Trust and Growth', *The Economic Journal*, **111**, 295–321.