

# Trust in the shadow of the courts if judges are no better

by

Geoffrey Brennan\*, Werner Güth\*\* and Hartmut Kliemt\*\*\*

RSS-Australian National University, Canberra\*      Humboldt University, Berlin\*\*

Gerhard Mercator University, Duisburg\*\*\*

## Abstract

Can a court system conceivably control opportunistic behavior if judges are selected from the same population as ordinary citizens and thus are no better than "the rest of us"? This paper provides a new and, as we claim, quite profound "rational choice" answer to that unsolved riddle. Adopting an indirect evolutionary approach with endogenous preference formation the complex interactions between "moral" intrinsic motivation to behave non-opportunistically and extrinsic "formal" controls of opportunism are analysed. Under the assumption that judges are no better than ordinary citizens it is shown that introducing a court system can nevertheless prevent that the more trustworthy are driven out. It cannot be excluded, though, that courts may themselves crowd out trustworthiness under certain circumstances.

Key words: Evolutionary game theory. Intrinsic motivation. Trust relationships. Court system. Legal litigation. Hobbesian problem of social order.

JEL Classification: A11, A13, C72, D74, K00, K12

## I. Introduction

Adopting an indirect evolutionary approach to human behavior the economist is no longer confined to an analysis of rational behavior relative to given preferences (and constraints). Individuals are still modelled as standard rational decision makers. But within an indirect

evolutionary approach the formation of the preferences underlying rational decision making can be treated as an endogenous aspect of a larger model in which the basic rational choice model is embedded. In the larger model it can be analysed how evolutionary success of individual "bearers" of preferences feeds back on the preferences underlying their rational decision making, and vice versa.

Subsequently we shall use such an indirect evolutionary approach to demonstrate that a court system sanctioning unfair behavior can work in favor of the more trustworthy even if judges are no better behaved than the population at large. Individuals who are serving as judges can be as opportunistic as the general population and yet introducing courts can eliminate the differential advantage of opportunists. But relying on behavioral control through extrinsic institutionalised incentives under specific parameter constellations can also favor the opportunists. It may "crowd out" intrinsic, non-opportunistic motivation.

The claim that there may be such negative side effects of "economising on love" (cf. on that Robertson 1956) through courts is directly opposed to the thesis that the courts will protect the less opportunistic. The claim has been frequently made ( cf. more recently in particular Frey 1997) but as far as we know it has not been discussed in a precise non-co-operative game model yet. The models provided subsequently do not only demonstrate that trust in the shadow of the courts may be warranted even if judges are no better than the population at large. They also spell out conditions under which the legal umbrella in fact may work to the disadvantage of the more trustworthy individuals.

More specifically, to understand the influence of "extrinsic control" on behavior we must know what behavior would have been like without "extrinsic motivation". Therefore the next section (II.1) recapitulates some previous results of ours about a basic "game of trust" in which extrinsic incentives to act "trustworthy" are lacking. Extrinsic incentives are "added" to the model, then, by introducing a "court system" (II.2). In the central part of the paper it is discussed whether and, if so, when

reliance on external enforcement may eliminate the advantages of opportunistic behavior even if judges are selected from the same population of players whose behavior they monitor. As already indicated, we do not dismiss the possibility, though, that introducing formal enforcement through courts may crowd out intrinsically motivated trustworthiness. As we shall show, if people behave non-trustworthy frequently anyway, then the provision of the extrinsic incentive of costly legal litigation may bring about a further decline of trustworthiness. On the other hand, if outside the shadow of the courts individuals behave non-trustworthy not too frequently, then introducing extrinsic incentives can in fact prevent that the more trustworthy are driven out. The general social climate or, more technically speaking, the population composition as well as relative enforcement costs are crucial for predicting the likely effects of extrinsic enforcement policies. But, as long as judges are not biased against those who behave fairly, there are always policies that can stabilise whatever degree of trustworthiness there is by means of a court system (III.). Finally, we put our discussion into a somewhat broader perspective (IV.).

## II. Basic modelling and background

### II.1. Some previous results

In our evolutionary model pairs of players of two types are drawn from a potentially infinite population and randomly matched to play a "*basic game of trust*". With equal probability each player of a pair is assigned the role of a first or that of a second mover respectively. In the role of the second mover the intrinsically motivated fair type will fairly reward a trustful first move of her co-player while the extrinsically motivated unfair type will respond by exploitation:

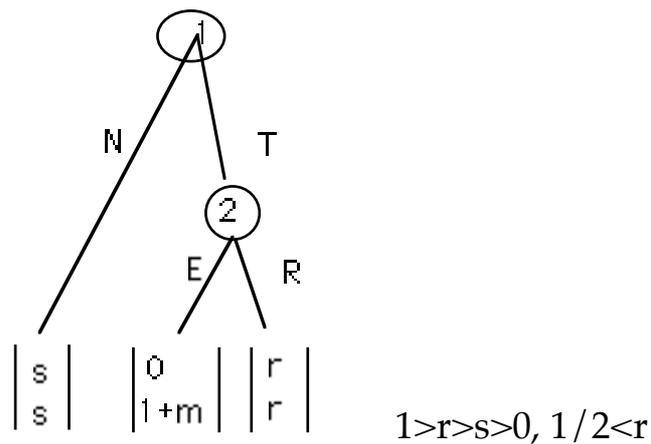


Figure 1

Player 1's is the first payoff while the second of the payoffs at the end nodes accrues to the second player. The parameters  $r, s, 0, 1$  are payoffs that refer to extrinsic incentives. They are based on some "objective" aspects of the real world like resources. These are directly related to evolutionary success. For convenience it is assumed that the subjective preferences over the objective parameter values can be represented by a scale with the same numerical payoffs.

The parameter  $m$  does not represent any external factor. It has no direct relationship to evolutionary success either. It is a subjective intrinsic motivational factor. For convenience we refer to it as the "conscience parameter" or more succinctly the "conscience" of the player. Since behavior of the second moving player is fully determined by the relationship between  $1+m$  and  $r$  we initially shall restrict attention to merely two parameter values of  $m$ :  $m=\bar{m}$ , with  $\bar{m} + 1 > r$ , and  $m=\underline{m}$ , with  $\underline{m}+1 < r$ . In short,  $\underline{m} < r-1 < \bar{m}$ . Malevolent intrinsic motivation is excluded by  $\bar{m} \leq 0$ . Later on, when discussing a variant of our model in which the behavior of the second mover is motivated in a more complex way we shall allow for continuous variation of the conscience parameter over the whole interval  $[\underline{m}, \bar{m}]$ .

Given the payoff structure described it should be obvious why we refer to the move  $N$  as a move of "no trust", to  $T$  as a move of "trust", to  $E$  as "exploitation" and to  $R$  as "reward". A rational player with  $m=\bar{m}$  who,

due to a lack of sufficient intrinsic motivation, will exploit a trusting first mover is referred to as being of the "non trustworthy" or "unfair  $\bar{m}$ -type". A player with  $m=\underline{m}$  who fairly rewards trust received is described as being of the "trustworthy" or "fair  $\underline{m}$ -type".

Putting the game of trust into an evolutionary perspective we assume that both types, in unbiased random matching, are drawn from a pool of players with a population composition characterised by the fraction  $p \in [0, 1]$  of trustworthy  $\underline{m}$ -types. The type composition parameter as well as the rules of the game are common knowledge among the players. Before matching the players do not know, though, whether they are going to play in the role of the first or in the role of the second mover. They are assigned their roles as first and second movers, respectively, by a chance move with equal probability for both outcomes.

We must distinguish two fundamentally different cases: complete and incomplete type information. Complete type information prevails if after matching their types are common knowledge among the players. Then a game of the following kind emerges:

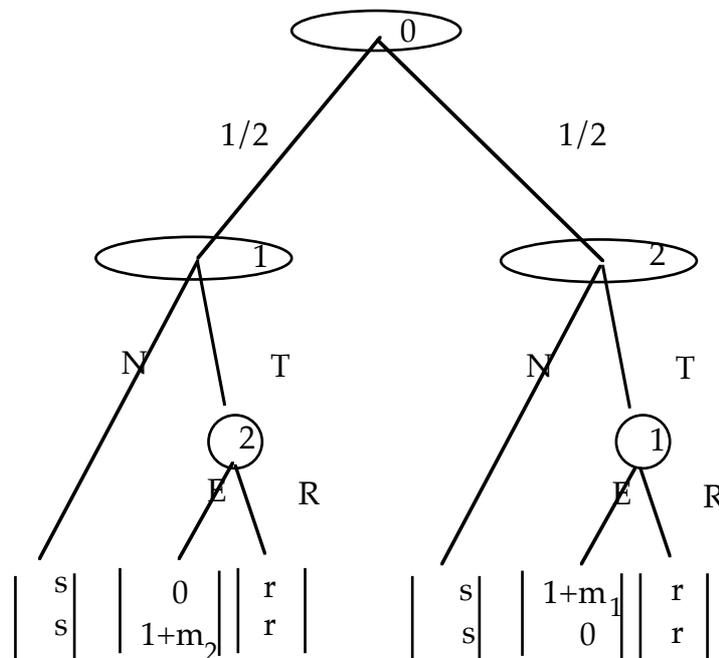


Figure 2

In this game it is commonly known among the players which type, either  $\underline{m}$  or  $\bar{m}$ , has been chosen for each of the players, that is they know  $m_i$ ,  $i=1, 2$ . Under this information condition it is intuitively plausible that being of the non-trustworthy type is a disadvantage. The unfair types will receive  $s$  whenever they end up in the role of the second mover. The fair types, however, will encounter trust whenever they move second and then receive  $r$ . Since  $r > s$  and since in the role of the first mover both types of players fare equal it is obvious that in the game of figure 2 only a population composition as characterised by  $p=1$  can be evolutionarily stable in the sense that there is a small neighborhood  $N_p$  of  $p$  such that for any population composition  $p' \in N_p$  the other type is less successful. Thus, once an evolutionarily stable population composition is reached it cannot be overthrown by an "invasion" of a small number of mutants. Since mutation is regarded as a small number phenomenon this captures a basic intuition of "stability".

In the case of incomplete type information the players remain ignorant of their respective types after matching. Collapsing the two chance moves determining the type of player 1 and of player 2 respectively into one, the remaining random event is the assignment of the roles of first and second mover. The following graph of figure 3 shows the basic characteristics of the game emerging then:

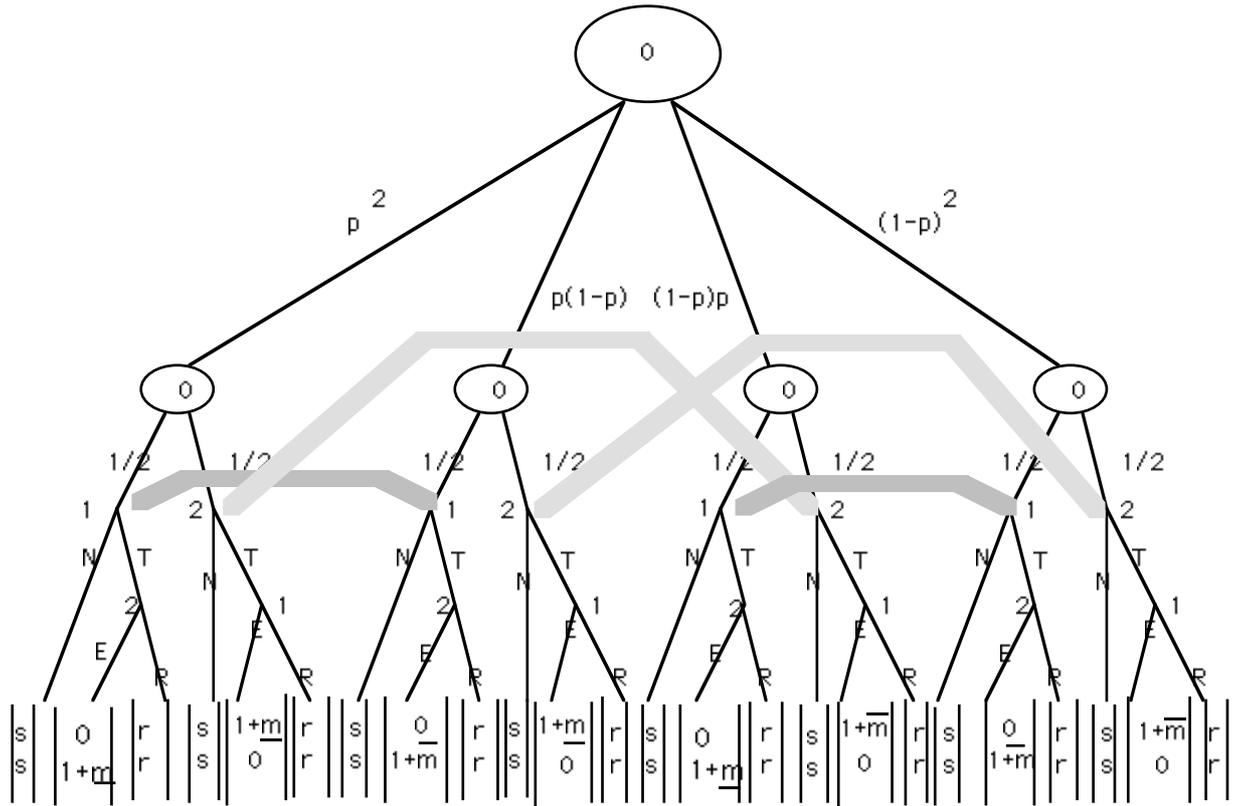


Figure 3

We have studied the evolutionary process based on this game in some detail elsewhere (cf. Güth and Kliemt 1994). Intuitively it should be quite obvious, though, what the likely outcome of the process is if types cannot be distinguished at all: After an initial choice of T, non-trustworthy types fare better than trustworthy ones in the role of the second mover. In that role the  $\bar{m}$ -types receive 1 as measured in "objective" payoff while the  $\underline{m}$ -types end up merely with  $r$ . Therefore, for all population compositions for which T is the rational choice of first movers the parameter  $p$  characterising the fraction of trustworthy types should decrease. For population compositions for which the rational choice of first movers is N both types indiscriminately receive  $s$ . We can differentiate between types, though, if we allow for a small minimum probability that players commit mistakes in executing their rational plans as required by the concept of a limit evolutionarily stable strategy or, for short, the LESS-concept (cf. Selten 1988). Then, with positive probability, occasional choices of T occur even though the rational plan

dictates to play N in the role of the first mover. If such "trembles" prevail the non-trustworthy types of players have an advantage over the trustworthy ones even for population compositions  $p$  for which showing trust, T, is not the first movers' rational choice. Only non-trustworthy types should be expected to survive -- at least eventually -- and only the  $\bar{m}$ -monomorphic population with  $p=0$  is evolutionarily stable according to the LESS-concept.

As long as types are completely indistinguishable and first movers choose T with some positive minimum probability, the non-trustworthy type should "drive out" the trustworthy one. Intuitively it should also be clear that this result might change if first movers can to some extent distinguish between second movers of different types before they decide between T and N. In fact, if a costly detection technology of limited reliability reveals information about the second mover's type *before* the first mover must make his initial move there may be two evolutionarily stable population compositions (cf. Güth and Kliemt 1995). Depending on the costs and the reliability of advance type detection we get either  $p=0$  or  $1 > p > 0$  (and in the limit of zero costs and full reliability even  $p=1$ ).

The following figure 4 sums up the previous results on the intermediate case in which some type information is available at some cost. It shall serve as a bench mark for our subsequent discussion of the influence of the courts on behavior and the population composition. The triangle in the graphic illustrates the population dynamics, i. e. the direction in which  $p$  moves in all likelihood, for alternative values of the parameter  $C'$  and alternative initial values of  $p$ .  $C'$  measures the costs of using a technology of given reliability for detecting trustworthy types before action. For each  $C$  one can calculate thresholds  $\underline{p}(C')$  and  $\bar{p}(C')$ . If  $p < \underline{p}(C')$ , then using the detection technology is too costly. Rational players will rather do without any specific type information than pay  $C'$  to acquire it. The game will basically be the one depicted in figure 3 and the share  $p$  of trustworthy individuals in the population will decrease until  $p=0$ . If  $\bar{p}(C') > p > \underline{p}(C')$  then usage of the detection technology is

worthwhile and will exert selective pressure in favor of the trustworthy until  $\bar{p}(C')$  is reached. For  $p > \bar{p}(C')$  exploitation occurs so rarely that players prefer to incur the remaining risk of being exploited to further reducing it at cost  $C'$ . Therefore, for  $p > \bar{p}(C')$  players will decide on not acquiring specific type information. They will thus play the game of figure 3 again. The population share  $p$  of trustworthy types will decrease until  $p = \bar{p}(C')$ .

As long as costs  $C$  are positive the non-trustworthy will never be driven out of the population completely. The population can become monomorphically trustworthy only if information is costless and thus the previously mentioned limiting case of complete type information emerges. We shall focus here on the more realistic and more central case of non-zero detection costs (cf. for another defence of the importance of that case Frank 1988). For  $C' \neq 0$  we assume that detection costs are sufficiently low and specific type information sufficiently reliable such that besides  $p=0$  also some  $p=\bar{p}(C') > 0$  can be stable.

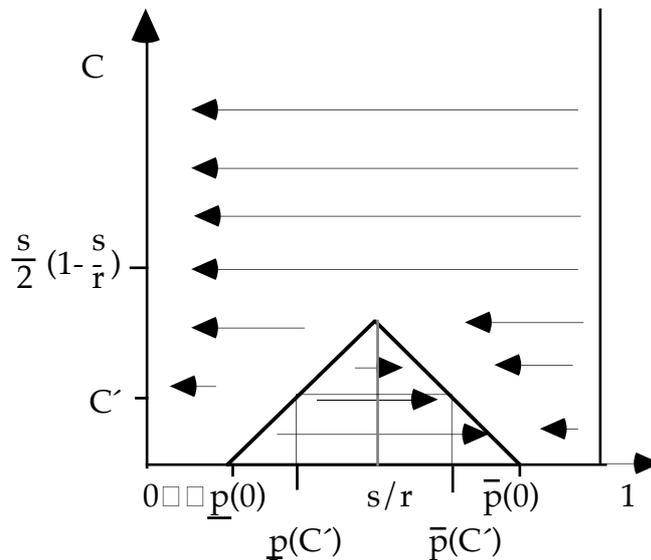


Figure 4

As remarked before the constellation described in figure 4 forms the bench mark to which we shall compare the situation emerges if the

courts cast their shadow on interactions like the basic game of trust. In the next step we shall introduce courts to our formal model.

### II.2. The basic model

In the model subsequently studied the costly information technology of limited reliability is substituted by a costly litigation procedure of limited reliability. Instead of the ability to detect their co-player's type before moving, first movers are able to rely on the external policing device after observing the outcome of the game. Players who have been exploited, can appeal to an "arbitrator" or "judge". More specifically, we assume that a game of the following form is played on each round of play.

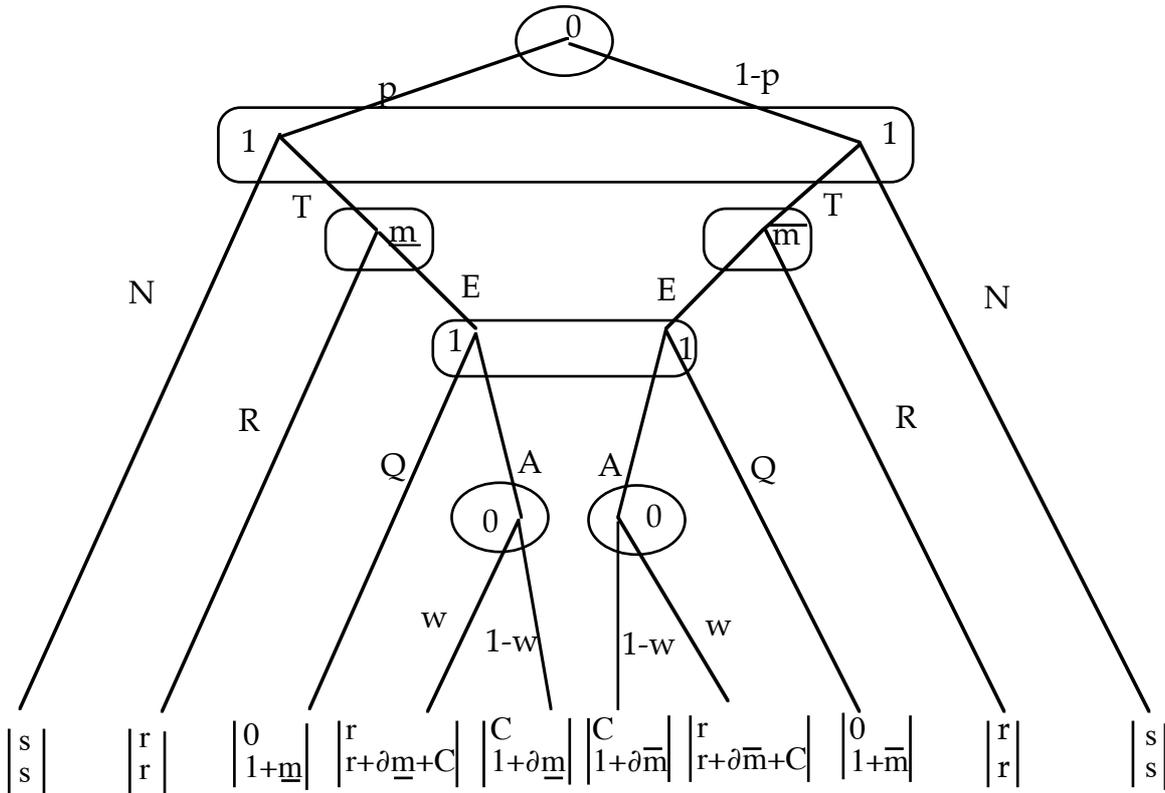


Figure 5

To keep things reasonably simple, figure 5 does not present the complete graph of the game. Since the first mover's type is irrelevant for the outcome of the basic game of trust we just present one of the subgames emerging after the roles of first and second mover have been assigned and the type of the first mover has been determined. In this subgame player 1 is assigned the role of the first and player 2 that of the second mover. The first mover knows this and his own type (which is irrelevant for his rational play anyway).

The subgame presented here captures the essential part of the interaction. It starts with nature's choice of the type of the first mover's co-player. Afterwards the first mover makes his initial move. After a non-trusting initial move, independently of the second mover's type, the result is the same as in the basic game of trust. After a trusting first move the second mover may decide on exploitation or on fair reward. In the simple game of trust this move would be determined completely by the second mover's type. The  $\underline{m}$ -type would fairly reward trust while the  $\bar{m}$ -type would choose to exploit the first mover's trust. However, now, the second mover anticipates that, after E, the first mover is going to move again. The first mover can either quit, Q, and thus give in to exploitation, or he can appeal, A, and thus try to reach restitution.

The consequences of each of these decisions depend on the decision of a "judge". "Nature" in a chance move chooses a judge from the same pool of individuals from which the players of the game of trust are drawn. With probability  $p$  the judge is a "trustworthy" or "fair" one. Such a fair judge upon appeal will decide on restitution if the second mover chose the exploitative alternative E. With probability  $1-p$  the judge will, however, be of the unfair type. An unfair judge with probability  $q$ ,

$$(II.1) \quad 0 < q < 1,$$

decides on restitution -- that is, she acts like a fair one -- and with complementary probability  $1-q$  decides against the plaintiff.

The judge controlling opportunistic behavior may seem like a "jack of the box". At first sight we are solving the problem of opportunism on the level of the game of trust by assuming it away on the level of norm-enforcement. However, in the model discussed here the judge is randomly drawn from the same population as the players. No selection effect is assumed to operate. Judges are not more trustworthy than non-judges. Neither would they necessarily play the role of the judge for more than one round of play. As far as these factors are concerned we do not make special assumptions about judges. We assume, though, that being a judge on any round of play is neutral with respect to evolutionary success of the players adopting that role. We can therefore treat the interaction as a two player game in which some of the outcomes depend on the type of the judge -- and thus on the population composition -- but not on any strategic choices affecting the judge's payoff.

Since, what judges do, can directly be derived from their type and the reliability parameter of the non-trustworthy type the parameters  $p$  and  $q$  are crucial for characterising judges' likely behavior. Since  $p$  and  $q$  are assumed to be commonly known, players can predict the likely consequences of appealing to a judge. The probability to encounter a non-trustworthy judge is just the same as that for being matched with a non-trustworthy co-player. However, since as assumed here  $q > 0$  applies, even the non-trustworthy judge may decide on restitution. This can be made plausible by the following line of argument: In their role of the judge the non-trustworthiness of players amounts to their unwillingness to invest sufficiently in the investigation of the case. Shirking on the judge's job they do reach the correct verdict only with reduced reliability.

The unreliable judge will not necessarily decide against restitution. With some probability  $q$  she will rather decide like the trustworthy  $\underline{m}$ -type of judge, that is, in favor of the exploited individual. Therefore the probability to reach restitution from a judge after being exploited is  $p$  --

the probability to encounter a trustworthy judge -- plus  $(1-p)q$  -- the probability to encounter a non-trustworthy judge who with probability  $q$  nevertheless decides on restitution. This explains

$$(II.2) \quad w = p + (1-p)q.$$

The latter, of course, implies  $1-w = (1-p)(1-q)$ .

The parameter  $\partial$  indicates how strongly extrinsic enforcement affects intrinsic motivation. In principle  $\partial$  could adopt any value in the closed interval  $[0, 1]$ . We shall discuss only the two pure or extreme cases in which  $\partial$  adopts one of the limits of the interval; that is

$$(II.3) \quad \partial \in \{0, 1\}.$$

If  $\partial=0$  prevails, going to court has a decisive effect on the second mover's perception of the situation. After the first mover engages her in a legal dispute she evaluates the situation exclusively according to "extrinsic standards". The first mover's second move of appealing to an external judge suspends all "inner" feelings of "guilt" she otherwise might have endorsed because of her exploitative behavior. The matter is "framed" as a purely legal dispute. This case will be studied in section III.1.

In view of the pervasiveness of "framing effects" (cf. Kahneman and Tversky 1984, Lindenberg 1983) the previous assumption is certainly plausible. Nevertheless, the second mover's feelings may be less affected by being sued. As indicated before, rather than considering all possible intermediate values we focus on the other extreme case that going to court does not at all affect the "inner feelings". This leads to the case  $\partial=1$  to be studied for a dichotomous type space in section III.2 and for a continuous type space in section III.3.

For our analysis we assume that  $\partial$  is common knowledge. Moreover, we presuppose that  $\partial$  is uniform across all interactions between all randomly matched pairs of players.

Finally, the variable  $C$  measuring the influence of the costs of using the legal procedure on individual utility is assumed to be negative:

$$(II.4) \quad C < 0.$$

### III. Trustworthiness in the shadow of the court

In the game of figure 5, initially, only the second mover knows her type. The first mover when deciding between  $T$  and  $N$  merely knows the population composition parameter  $p$ . However, at his second information set, the first mover knows that  $T$  and  $E$  must have been played. Conceivably he could infer the type of the second mover from this information. This would be the case if only the  $\bar{m}$ -type would choose  $E$  in the role of a second mover. Then the first mover when having to decide between  $Q$  and  $A$  would know that with probability 0 he would be at the left and with probability 1 at the right decision node.

Since prospects are completely identical for player 1 at both nodes of his second information set the subjective probability of being at the right or at the left node is irrelevant for rational play, however. We can therefore disregard the signalling aspects of the game of figure 5. The expectation after  $A$  is  $(1-w)C + wr$  while after  $Q$  it is 0, regardless. We need to consider only the relation between  $(1-w)C + wr$  and 0.

If  $(1-w)C + wr < 0$  applies player 1 does not plan to appeal. Since this is common knowledge among the players the dominated alternative,  $A$ , of appealing can be deleted from figure 5. The game of figure 6 emerges. As can easily be checked this game corresponds to one of the four information sets for the first mover in figure 3. Thus, the complete tree of this game would be equivalent to the game without external enforcement as presented in figure 3 above. Therefore all the results sketched in our intuitive discussion of that game carry over to the case of no appeal.

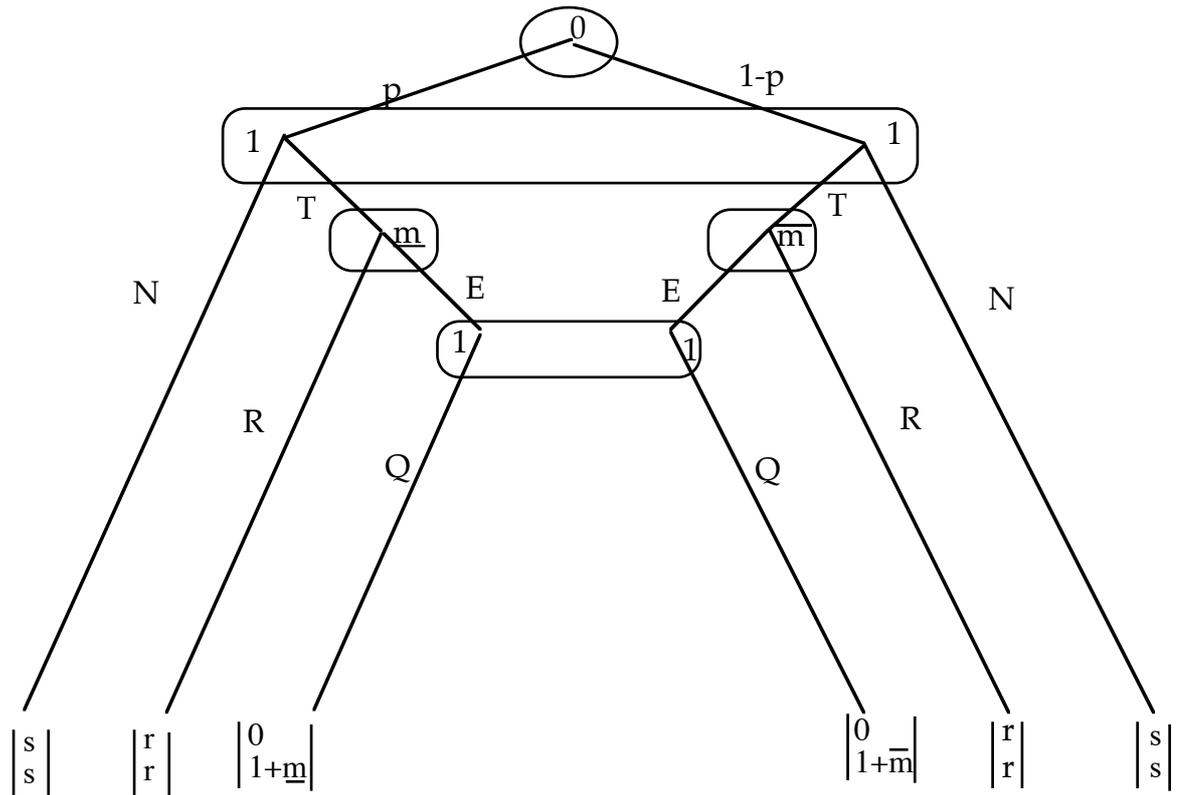


Figure 6

Subsequently the main focus is on parameter constellations under which the external enforcement mechanism is in fact used. In dealing with the likely dynamics and stable values of  $p$  under various constellations we shall distinguish between the case  $\partial=0$ , appealing to the court suppresses feelings of "guilt" completely, and the case  $\partial=1$  in which the extrinsic motivational effects of court intervention are merely superimposed on the pre-existing motivational structure. Intermediate values  $\partial \in (0, 1)$  could be studied along the same lines. But, since this would not lead to deeper additional insights, it suffices to deal with merely two extreme values of  $\partial$ .

Rather than studying the continuous variation of  $\partial$  over the whole interval  $[0, 1]$  we shall analyse a continuous variation of  $m$  over the whole interval  $[\underline{m}, \bar{m}]$ . We shall refer to this, as the "continuous" case as opposed to the "dichotomous" case with mutant space  $\{\underline{m}, \bar{m}\}$ . In fact, we use the distinction between these two cases to introduce a more

fundamental further one. The two cases differ not merely with respect to their mutant spaces. As discussed here, they lead to two different models which, though formally similar, are based on fundamentally different concepts of trustworthiness. In the dichotomous case trustworthiness is a "motivational" concept. Individuals are trustworthy or they are not, pending on the strength  $|m|$  of their conscience. In the continuous case trustworthiness becomes a "behavioral" rather than a motivational concept. Individuals are deemed trustworthy or non-trustworthy according to their expected rational choices which depend not only on the strength of their conscience but also on the population composition.

### III.1. Dichotomous case $\partial=0$

In the preceding comments on figure 6 we have dealt already with  $(1-w)C + wr < 0$ . Assume therefore

$$(III.1) \quad (1-w)C + wr > 0.$$

For all  $p$  fulfilling this relation the first mover plans on choosing A. Though, conceivably, the value of the parameter  $m$  might differentially affect strategic choices it does not in the present case. Unless the condition for appeal is reversed both types of players face exactly the same strategic incentives if assigned the role of the second mover. For, after being sued, the second mover regardless of her type "frames" the interaction as a purely legal matter.

Thus, if  $\partial=0$ , the anticipated choice of A suppresses all intrinsic motivational factors like "guilt" or a "bad conscience". Moreover, the parameter  $m$  is not directly related to evolutionary success. It can be left out of account in calculating the contribution of the results of a play to success in evolutionary competition. Therefore, since type or intrinsic motivation does not affect choice if  $\partial=0$  prevails and A is chosen, evolutionary success of both types of players must be the same. Finally,

payoff in the role of the arbitrator is independent of the type of the player as well.

In sum, if in case  $\partial=0$  condition III.1 applies, we cannot differentiate between  $\underline{m}$ -types and  $\bar{m}$ -types with respect to their likely evolutionary success. Any population composition parameter  $p$  implying appeal can be weakly stable since for all  $p \in [0, 1]$  there exists a generic neighborhood  $N_p$  of  $p$  relative to  $[0, 1]$  such that for  $p' \in N_p$  all types are equally successful.

This argument does not change if players in the role of the second mover anticipate that exploited first movers once in a while fail to appeal. For, if the first mover appeals, then in all generic cases, both types of a second mover face the same incentives and thus behave the same way. For sufficiently small trembles leading to a choice of T either both types strictly prefer R or both types strictly prefer E. Therefore, if players in the role of the first mover have a sufficient incentive to choose A no systematic differentiation between types can take place and thus no crowding out effect can emerge.

In case of population composition parameters  $p$  for which  $Q$  is rational, the game of figure 6 is played. As we know from our above discussion the  $\bar{m}$ -monomorphic population characterised by  $p=0$  is the only stable one in this game.

The following figure sums up the essentials of this discussion of the stability and dynamics of  $p$  in the dichotomous case  $\partial=0$ :

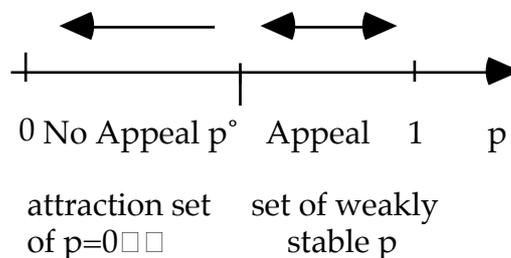


Figure 7

The critical value  $p^\circ$  separating the region of "no appeal", Q, from that of appeal, A, can be derived by solving  $(1-w)C + wr=0$  for  $p$ ,

$$(1-w)C + wr=0$$

$$\Leftrightarrow p = \frac{-qr - C(1-q)}{(1-q)(r-C)},$$

yielding

$$(III.2) \quad p^\circ := \frac{qr + C(1-q)}{(1-q)(C-r)}.$$

The first mover should appeal, A, if  $p > p^\circ = \frac{qr + C(1-q)}{(1-q)(C-r)}$ , and he should not appeal, Q, if  $p < p^\circ = \frac{qr + C(1-q)}{(1-q)(C-r)}$ .

Note  $p^\circ = \frac{qr + C(1-q)}{(1-q)(C-r)} = \frac{-q}{r-C} \frac{r - C}{1-q}$ . Thus,  $p^\circ$ , is negative if  $\frac{q}{1-q} r > -C$ . Of course, if  $p^\circ < 0$  then  $p > p^\circ$  for all  $p$ . First movers will appeal for all population compositions  $p$  rendering them weakly stable. Thus, the set of weakly stable strategies can comprise the whole interval  $[0, 1]$  in figure 7.

On the other hand,  $p^\circ$  is always smaller than 1. Therefore the attraction set of  $p=0$  corresponding to a monomorphic population composition of unfair  $\bar{m}$ -types can never be exhaustive. Thus, even though only  $p=0$  can be stable in the sense of the LESS-concept if  $\frac{q}{1-q} r < -C$ , there are always non-monomorphic population compositions that are weakly stable.

If  $p \in (p^\circ, 1)$  and thus first movers after being exploited appeal then the courts play a role. After play they have the relevant type information and use it to the advantage of the trustworthy. Though the courts cannot "crowd in" trustworthy types they can keep them from being crowded out. Recalling our bench mark case with the "old" detection costs  $C'$  this is particularly relevant for  $p > \bar{p}(C') \geq p^\circ$ . In this realm  $p$  can be rendered

weakly stable by court intervention. On the other hand, the fact that merely  $p=0$  can be (limit)evolutionarily stable is not due to a crowding out effect of the relevant kind, quite to the contrary. If the trustworthy loose out for  $p \in (\bar{p}(C'), p^\circ)$  they are *not* crowded out *by* the courts. As the previous discussion of the game depicted in figure 3 showed the trustworthy would have been eliminated anyway under the prevailing information conditions. And, this lack of information is not the courts' fault, so to say. Let us see whether these results change for  $\partial=1$ .

### III.2 Dichotomous case $\partial=1$

Consider again the game of figure 5. However, now assume  $\partial=1$ . As before, player 1 does never intentionally choose to appeal iff  $p < p^\circ$  (or  $wr + (1-w)C < 0$ ). After deleting the respective moves, the game is reduced to that of figure 6 to which our previous results apply. It suffices to analyse the game under the assumption  $p > p^\circ$  implying that first movers appeal after exploitation.

#### III.2.1. Rational play

Note first, that the parameter  $m$  does not affect the first moving player's second decision. If, as assumed here,  $p > p^\circ$  then the first mover independently of his type appeals. This must be taken into account by the second mover in her choice between fair reward,  $R$ , and exploitation,  $E$ . The payoff expectation after  $R$  is  $r$  while after  $E$  it is  $w(r+m+C) + (1-w)(1+m)$ . The second mover prefers  $E$  over  $R$  iff

$$w(r+m+C) + (1-w)(1+m) > r$$

$$(III.3) \quad m > (r-1)(1-q) - Cq + p(1-q)(1-r-C)$$

Let  $\alpha := (r-1)(1-q) - Cq$  and  $\beta := (1-q)(1-r-C)$  and  $\omega(p) := \alpha + \beta p$ .

Since  $0 \leq q < 1$  and  $1 > r > 0 > C$ , obviously  $\beta > 0$ . Therefore  $\omega$  is a positive (affine) linear function of  $p$ . It can be used to reformulate III.3 more succinctly yielding:

The second mover prefers E over R iff

$$(III.4) \quad m > \omega(p).$$

Note that  $1 + \underline{m} < r$  and  $C < 0$  imply  $w(r + \underline{m} + C) + (1-w)(1 + \underline{m}) < r$ . Thus, for the trustworthy  $\underline{m}$ -type the condition  $\underline{m} < \omega(p)$  holds good for all population compositions. For the dichotomous case we have  $m \in \{\underline{m}, \bar{m}\}$ , with  $\underline{m} < r - 1 < \bar{m}$  and can therefore describe behavior of the two types of second movers in the following way:

The trustworthy  $\underline{m}$ -type chooses fair reward, R, for all population compositions.

The non-trustworthy  $\bar{m}$ -type chooses fair reward, R, if  $\bar{m} < \omega(p)$ .

The non-trustworthy  $\bar{m}$ -type chooses E if  $\bar{m} > \omega(p)$ .

In short, if  $p > p^\circ$ , and thus the courts are used, then exploitative behavior E will be observed only if  $\bar{m} > \omega(p)$ .

Let us finally turn to the first move of the first moving player. His decisions to trust, T, or not to trust, N, are type-independent. If the condition for appeal,  $p > p^\circ$ , is fulfilled N can be rational only if  $\bar{m} > \omega(p)$ . For, if  $\bar{m} < \omega(p)$  applies, T is always rewarded by R and thus rational. Provided that  $p > p^\circ$  applies we have:

$$(i) \quad \bar{m} < \omega(p)$$

$\Rightarrow$  the first mover prefers T over N.

$$(ii) \quad \bar{m} > \omega(p) \text{ and } pr + (1-p)((1-w)C + wr) > s$$

$\Rightarrow$  the first mover prefers T over N.

(iii)  $\bar{m} > \omega(p)$  and  $pr + (1-p)((1-w)C + wr) < s$

$\Rightarrow$  the first mover prefers N over T.

Analysing the second condition for preferring N over T stated in (iii) yields

$$p(r) + (1-p)((1-w)C + wr) < s$$

$$\Leftrightarrow p(2-p) < \frac{\frac{s-C}{r-C} - q}{1-q} \Leftrightarrow$$

$$(III.5) \quad p(2-p) < \eta;$$

$$\text{where } \eta := \frac{\frac{s-C}{r-C} - q}{1-q}.$$

Clearly,  $1 > \frac{s-C}{r-C} > 0$  and thus  $\eta < 1$ . Moreover  $\eta < 0$  iff  $q > \frac{s-C}{r-C}$ . Thus, independently of the population composition T is chosen if  $q > \frac{s-C}{r-C}$ . Note also that  $p(2-p)$  is monotonically increasing over the interval  $(0, 1)$ . The threshold beyond which the condition for trust rather than no trust is fulfilled is  $p = 1 - [1 - \eta]^{1/2}$  since  $p(2-p) = \eta \Leftrightarrow p = 1 \pm [1 - \eta]^{1/2}$ . Thus the second condition for showing non-trustful behavior as stated in (iii) is fulfilled for

$$(III.6) \quad 0 \leq p < 1 - [1 - \eta]^{1/2}.$$

Let us summarise:

For  $\eta < 0$                       T is chosen for all p;

For  $\eta \geq 0$

$p > 1 - [1 - \eta]^{1/2} \Rightarrow$  T is chosen;

$p < 1 - [1 - \eta]^{1/2} \Rightarrow$  N is chosen if  $\bar{m} > \omega(p)$ .

We can now fully describe solution play in case  $\partial=1$ . Basically, if  $p > p^\circ$  we have to deal with the following cases (where differentiation between types concerns only play in the role of the second mover since types behave identically in the role of the first mover):

#### Case 1

If  $\eta < 0$  or  $p > 1 - [1 - \eta]^{1/2}$

then for all population compositions fulfilling  $\bar{m} > \omega(p)$

solution play is (T, A; E) for  $m = \bar{m}$  and (T, A; R) for  $m = \underline{m}$ .

#### Case 2

If  $\eta > 0$  and  $0 \leq p < 1 - [1 - \eta]^{1/2}$

then for all population compositions fulfilling  $\bar{m} > \omega(p)$

solution play is (N, A; E) for  $m = \bar{m}$  and (N, A; R) for  $m = \underline{m}$ .

#### Case 3

For all population compositions fulfilling  $\bar{m} < \omega(p)$

solution play is (T, A; R) for both m-types.

### III.2.2. Evolutionary dynamics and stability

Only different behavior in the role of the second mover can lead to differential evolutionary success of different types. Otherwise types behave and fare the same way. Therefore, in discussing evolutionary

dynamics and stability, it suffices to focus on behavior in the role of the second mover.

Moreover, as stated before the game of figure 6 emerges if  $p < p^\circ$ . Second movers behave according to type then. Therefore,  $p$  decreases for population compositions  $p$  with  $p < p^\circ$  as long as there is a positive probability for occasional choices of T. In short, if the courts are *not* used,  $p$  will tend to decrease. Clearly, this does not amount to crowding out *by* the courts. But the courts not being used do not enhance the survival prospects of the more trustworthy either.

Only if  $p > p^\circ$  the courts are used and can conceivably have some impact on survival prospects. Neglecting the relationship between  $p$  and  $\underline{p}(C')$  as well as  $\bar{p}(C')$  for the time being we may restrict attention to differential success of different types' play in the role of the second mover when  $p > p^\circ$ . On the one hand, there are population compositions  $p$  for which both types of players in view of  $\bar{m} < \omega(p)$  choose R. Again, since both types behave the same way, nothing can differentiate between them. On the other hand, there are population compositions with  $\bar{m} > \omega(p)$  for which the two types of players behave differently in the role of the second mover. Since they behave differently types can differ in evolutionary success then. -- To facilitate orientation let us first give a kind of bird's eye view of the argument.

Recall, that  $\omega$  is strictly monotonic. Therefore  $\omega^{-1}$  exists yielding a unique value  $\omega^{-1}(m)$  for all  $m$ . One can thus define

$$\bar{p} := \omega^{-1}(\bar{m}).$$

Using III.3 yields

$$\bar{p} = \frac{\bar{m} + (1-r)(1-q)\square + \square qC}{(1-q)(1-r-C)}.$$

For  $p > \bar{p}$  the  $\bar{m}$ -type second movers will choose R rather than E and thus behave the same way as the  $\underline{m}$ -types in the role of the second mover. For

$p < \bar{p}$  the  $\bar{m}$ -type second movers will choose E rather than R and thus behave differently from  $\underline{m}$ -type second movers.

The following graphical illustrations of population dynamics for  $p \in [0, p^\circ)$  or  $p \in (p^\circ, 1]$  -- Figure 8a -- and  $p \in (p^\circ, \bar{p})$  or  $p \in (\bar{p}, 1]$  -- Figure 8b -- give an overview over claims to be justified in the subsequent analytical discussion of cases and sub-cases.

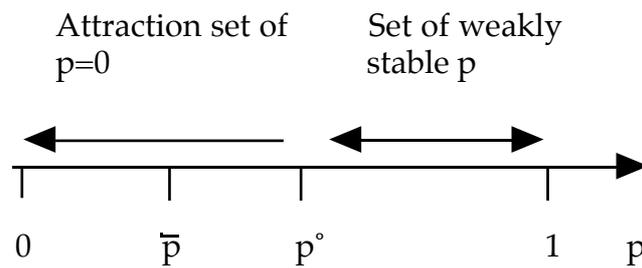


Figure 8a

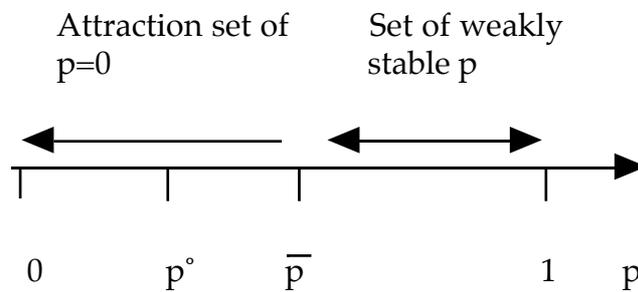


Figure 8b

Figure 8a

Starting with figure 8a, i.e. with the case  $p^\circ > \bar{p}$ , we must consider  $p \in [0, p^\circ)$  as well as  $p \in (p^\circ, 1]$ . As noted before, if  $p < p^\circ$  this leads back to the game of figure 6 (leaving out the relation to  $\bar{p}(C')$ ,  $\underline{p}(C')$  completely). We can state therefore: If  $p \in (0, p^\circ)$  then  $p$  decreases.

Obviously, the parameter constellation  $p^\circ > \bar{p}$  depicted in figure 8a is such that  $p^\circ < p$  implies  $p > \bar{p}$ . We therefore have  $p > p^\circ$  and  $p > \bar{p}$ . In this case

$\bar{m} < \omega(p)$  and thus both types act fairly if assigned the role of the second mover. All players trust and their trust is met by fair reward. Fair and unfair types behave the same way in the role of the second mover. Thus, for population compositions  $p \in (p^\circ, 1]$  -- i. e.  $p$  satisfying  $p > p^\circ$  and  $p > \bar{p}$  -- there will neither be a systematic increase nor a systematic decrease of  $p$ .

### Figure 8b

For parameter constellations like those depicted in figure 8b where  $p^\circ < \bar{p}$  we must consider  $p \in (p^\circ, \bar{p})$  as well as  $p \in (\bar{p}, 1]$ . The former amounts to  $p > p^\circ$  and  $p < \bar{p}$  and the latter to  $p > p^\circ$  and  $p > \bar{p}$ .

$p > p^\circ$  and  $p > \bar{p}$ :

This case has been dealt with already in the discussion of figure 8a. We know that under these conditions  $p$  neither systematically increases nor systematically decreases. It may be noted in passing, though, that  $\omega(0) = \alpha + \beta 0 = (r-1)(1-q) - qC > \bar{m}$  implies  $\omega(p) > \bar{m}$  or  $p > \bar{p}$  for all  $p$ , since  $\omega$  is monotonically increasing in  $p$ .

$p > p^\circ$  and  $p < \bar{p}$ :

The premise  $p < \bar{p}$  is equivalent to  $\bar{m} > \omega(p)$  which in turn is equivalent to  $w(r + \bar{m} + C) + (1-w)(1 + \bar{m}) > r$ . Recall we assumed  $\bar{m} \leq 0$ ,  $C < 0$ . Therefore  $\bar{m} > \omega(p)$  implies that the payoff components directly related to evolutionary success fulfil  $w(r + C) + (1-w)(1) > r$ . The  $\bar{m}$ -type's "objective" payoff,  $w(r + C) + (1-w)(1)$ , is larger than the "objective" payoff,  $r$ , of the  $\underline{m}$ -type. Consequently, the population share,  $p$ , of trustworthy  $\underline{m}$ -types decreases. For  $\eta < 0$ ,  $T$  is chosen and the argument applies as long as  $p > p^\circ$ . For  $\eta > 0$  the population share  $p$  of trustworthy  $\underline{m}$ -types declines until  $p < 1 - [1 - \eta]^{1/2}$ . The very moment this threshold is reached case 2 of the above characterisation of solution play emerges and the ESS

concept itself does not further differentiate between types. But, under the assumption that  $T$  is chosen at least as an occasional chance move with small but positive probability, case 2 basically can be treated like case 1. Since the two types behave differently if  $\bar{m} > \omega(p)$ , an occasional move  $T$  may be expected to bring about differential evolutionary success. Therefore  $p$  should decrease according to the LESS-concept. We can thus state: For any population composition  $p \in (p^\circ, \bar{p})$  the share  $p$  of trustworthy types decreases until  $p = p^\circ$ . As the discussion of figure 8a shows  $p = p^\circ$  is merely transient and  $p$  will decline beyond -- again the general disclaimer that the possible interference of the costly detection technology is left out of account, of course, applies.

The upshot of the preceding discussion is that both,  $p < \bar{p}$  as well as  $p < p^\circ$ , are sufficient for bringing about a decrease of  $p$ . Since either will do, the maximum of the two sufficient conditions is all that matters:

For  $p < \max\{\bar{p}, p^\circ\}$  types behave differently and the share of trustworthy types decreases.

For  $p > \max\{\bar{p}, p^\circ\}$  types behave the same way and the population composition is weakly stable.

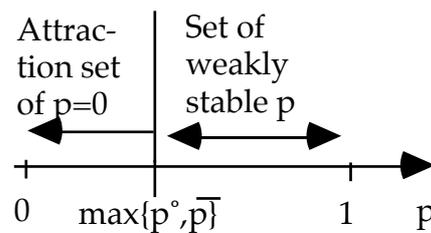


Figure 9

In view of figure 9 the following observations about evolutionary stability are obvious:

$\max\{\bar{p}, p^\circ\} > 0$  then the population composition characterised by  $p = 0$  is stable in the sense of the LESS-concept.

$p > \max \{ \bar{p}, p^\circ \}$  then the population composition characterised by  $p$  is weakly stable.

Substituting the definitions of the terms we get:

$$\max \left\{ \frac{\bar{m} + (1-r)(1-q) + qCr + C(1-q)}{(1-q)(1-r-C)}, \frac{qCr + C(1-q)}{(1-q)(C-r)} \right\} > 0 \Rightarrow$$

$p=0$  is stable (in the LESS sense)

and

all  $p > \max \left\{ \frac{\bar{m} + (1-r)(1-q) + qCr + C(1-q)}{(1-q)(1-r-C)}, \frac{qCr + C(1-q)}{(1-q)(C-r)} \right\}$  are weakly stable;

$$\max \left\{ \frac{\bar{m} + (1-r)(1-q) + qCr + C(1-q)}{(1-q)(1-r-C)}, \frac{qCr + C(1-q)}{(1-q)(C-r)} \right\} < 0 \Rightarrow$$

all population compositions are weakly stable.

The result of our discussion of the dichotomous case  $\partial=1$  is the following: Given certain parameter constellations, the shadow of the courts can counteract the tendency towards eliminating the trustworthy from the basic evolutionary game of trust. For this result to emerge it is not necessary that judges are more trustworthy than the population at large. A generic attraction set of  $p=0$  may emerge for certain values of punishment  $C$ , though. But for  $p < p^\circ$  the decrease in trustworthy individuals is definitely not brought about *by* the courts. Rather a general lack of intrinsic motivation in society induces the "decline of morals" since it also affects judges' behavior. A crowding out effect in the proper sense of that term could be conceivably observed only for  $p \in (p^\circ, \bar{p}) \neq \emptyset$ . Only for  $p$  from the interval  $(p^\circ, \bar{p})$  first movers do in fact appeal to the courts after being exploited. But crowding out also

presupposes that a non-monomorphic population of trustworthy and non-trustworthy individuals with a limit evolutionarily stable population composition parameter  $p \in (p^\circ, \bar{p})$  and  $p = \bar{p}(C') > 0$  has emerged (cf. on this the discussion of figure 4 above). Only if this applies can we say that the trustworthy are driven out through the influence of the courts.

According to the previous line of argument, the possibility that crowding out may emerge must be taken into account when considering policy measures like that of imposing a court system on a realm of human interaction previously unregulated. However, an exclusive focus on this mere possibility would lead to a very distorted view of possible policy measures. Rejecting legal enforcement by a court system simply because it may conceivably have negative side effects on trustworthiness is unwise. The possible positive effects of courts must not be neglected. -- As we shall see next, if judges in their role are not biased in favor of the non-trustworthy, then introducing extrinsic enforcement can render any population composition weakly stable provided that the costs of legal litigation are suitably set.

### III.2.3. When the courts can prevent elimination of the trustworthy

As noted before, if  $\max \{ \bar{p}, p^\circ \} < 0$ , then all population compositions are weakly stable. Moreover, if  $\max \{ \bar{p}, p^\circ \} < 0$ , then, as far as overt behavior is concerned, we live in a "morally perfect world". Under these conditions all actors independently of their type and thus of the strength of their conscience rationally trust and intend to retribute fairly while after and occasional mistaken unfair move of second movers, as first movers they plan on appealing to the courts. In short, all players plan to choose according to (T, A; R). In such situations the introduction of courts as such can guarantee that all players are well behaved even though there is no selection effect operating in favor of judges who behave better than "the rest of us". Clearly, from a policy point of view, measures to bring

about this situation deserve special attention. So let us study possible such measures in some detail.

Both,  $\bar{p}$  as well as  $p^\circ$  depend on the cost parameter  $C$ . At the same time  $C$  is the crucial policy variable when considering the introduction of a court enforcement system. At least conceivably, policy makers should be able to choose it as seems fit. This raises the question under what circumstances there is an interval from which  $C$  can be chosen such that  $\max\{\bar{p}, p^\circ\} < 0$ .

Now,

$$p^\circ < 0 \quad \Leftrightarrow \quad \frac{qr + C(1-q)}{(1-q)(C-r)} \stackrel{-q}{=} \frac{1-q}{r-C} < 0 \quad \Leftrightarrow \quad \frac{q}{1-q} r > -C;$$

then for all population compositions  $A$  is chosen.

$$\bar{p} < 0 \quad \Leftrightarrow \quad \frac{\bar{m} + (1-r)(1-q)}{(1-q)(1-r-C)} < 0 \quad \Leftrightarrow \quad -C > \frac{\bar{m} + (1-r)(1-q)}{q};$$

then for all population compositions  $R$  is chosen which, of course, implies that the choice of  $T$  is rational for all population compositions as well.

The interval  $(\frac{\bar{m} + (1-r)(1-q)}{q}, \frac{q}{1-q} r)$  of cost parameters guaranteeing  $(T, A; R)$  to be rational is non-empty iff

$$\begin{aligned} & \frac{q}{1-q} r > \frac{\bar{m} + (1-r)(1-q)}{q} \\ \Leftrightarrow & \quad r > \frac{\bar{m}}{q^2} (1-r) \left(\frac{1-q}{q}\right)^2, \end{aligned}$$

which is fulfilled in any case for  $r, q > \frac{1}{2}$ . Thus, if  $r, q > \frac{1}{2}$ , there are  $C$  such that  $-C \in (\frac{\bar{m} + (1-r)(1-q)}{q}, \frac{q}{1-q} r)$ . In figure 1 above  $r > \frac{1}{2}$  was already imposed. As in the parallel case of analyses of the standard prisoner's dilemma we required for the basic game of trust, that it should not be

better for the players to take turns in being exploited rather than playing fair twice. Thus the crucial additional requirement is  $q > \frac{1}{2}$ . In fact,  $q \geq \frac{1}{2}$  does as well. Therefore, whenever shirking of judges would mean that they simply use an unbiased random mechanism there would always be cost parameters  $C$  such that the court system exerts no evolutionary drive at all while inducing fair behavior throughout.

A wise regulator intending "to make the world a better place" would want to choose such values of  $C$ . Doing this, she would also be assured that the court system would not exert a crowding out effect itself. As far as the regulator's range of choices is concerned, much depends, though, on what shirking of non-trustworthy judges amounts to. If there is merely a slightly reduced reliability  $q$ , which is close to 1, say  $q = 1 - \varepsilon$  with  $\frac{1}{2} \gg \varepsilon > 0$ , then practically all  $C$  will do. For,  $\frac{\bar{m} + (1-r)(1-1+\varepsilon)}{1-\varepsilon} \rightarrow \bar{m} \leq 0$  if  $\varepsilon \rightarrow 0$ , while  $\frac{1-\varepsilon}{1-1+\varepsilon} r$  grows indefinitely if  $\varepsilon \rightarrow 0$ .

As long as we can sufficiently trust individuals when adopting the role of a judge, court intervention will affect behavior, and only behavior. For  $q$  close to 1, any population composition can be weakly stable if  $C$  is appropriately chosen. The latter is not a big "if" for  $q$  close to 1, however.

A very high value of the parameter  $q$  could also be interpreted as expressing the fact that less of a conscience is necessary to overcome temptations in the role of a judge than in the role of an ordinary player. There is a threshold other than  $r-1$  telling apart unreliable and reliable judges then. In particular, rewards and temptations directly related to performance may be systematically lowered by tenured positions of judges, fixed salaries etc. Such asymmetries in costs and systematic differences of behavior in low cost as opposed to high cost situations might be exploited in policies influencing  $q$ . One may also have some second thoughts about how decision rules for juries functioning as judges could influence  $q$  and, perhaps, improve performance. However, such considerations go against the grain of the present argument which,

of course, tries to avoid any assumption that judges behave differently from the rest of the population and yet courts can improve performance.

Let us note, though, that the story changes fundamentally if the shirking of judges leads to  $q < 1/2$ . In particular, if  $q \ll 1/2$  then the non-trustworthy judges are strongly biased towards the exploiters. If  $q$  becomes sufficiently small the interval for  $C$  becomes empty. Not surprisingly, with such judges there cannot be a court system stabilising universally fair behavior and eliminating the evolutionary advantage of the unfair types in a world in which type information is lacking. Of course, as we have seen the opposite is true as well. If sufficiently fair judges for some exogenous reason or other exist, initially, then one can stabilise co-operative behavior throughout and thus stop any evolutionary pressure favoring the non-trustworthy by introducing a court system with a suitably chosen cost parameter.

It may finally be noted that the previous argument applies for all  $\bar{m} \leq 0$ . We could thus make the following claim: As long as judges are not biased towards the unfair -- that is, as long as  $q \geq 1/2$  -- and as long as players are not intrinsically malevolent -- that is, as long as  $\bar{m} \leq 0$  -- one can always find suitable  $C$  such that all population compositions are weakly stable and rational play universally conforms with  $(T, A; R)$ .

To see this in some detail, note

$$\omega(0) = (r-1)(1-q) - qC > 0 \Leftrightarrow -C > \frac{(1-q)(1-r)}{q}.$$

The interval  $(\frac{(1-q)(1-r)}{q}, \frac{q}{1-q}r)$  of cost parameters  $C$  for which reward,  $R$ , and appeal,  $A$ , can be guaranteed to be the rational choice independently of both  $m$  and  $p$  and thus for all conceivable types and population compositions is non-empty iff

$$\frac{q}{1-q}r > \frac{(1-q)(1-r)}{q} \Leftrightarrow r > (1-r)\left(\frac{1-q}{q}\right)^2,$$

which is again fulfilled for  $r, q > \frac{1}{2}$ . In fact, again it suffices if for at least one of the two parameters  $r, q$  the strict inequality holds. Relying on the above assumption of  $r > 1/2$  again, we can allow for  $q \geq 1/2$  and thus for an unbiased random choice on the judges' side.

Since  $\bar{m} \leq 0$ , we have  $r > (1-r) \left(\frac{1-q}{q}\right)^2 > \frac{\bar{m}}{q^2} (1-r) \left(\frac{1-q}{q}\right)^2$  or, to put it slightly otherwise,  $\left(\frac{(1-q)(1-r)}{q}, \frac{q}{1-q} r\right) \subseteq \left(\frac{\bar{m} + (1-r)(1-q)}{q}, \frac{q}{1-q} r\right)$ .

Since for  $C$  with  $-C \in \left(\frac{\bar{m} + (1-r)(1-q)}{q}, \frac{q}{1-q} r\right)$  all  $p$  are weakly stable, this demonstrates the claim raised above.

In the final discussion we shall come back to the preceding considerations. But, as announced in our introductory remarks, let us turn first to the behavioral as opposed to the motivational interpretation of trustworthiness. This interpretation suggests a fundamentally different outlook on the problems so far discussed. Thus, our subsequent discussion of the case of a continuous variation of  $m$  is not merely some extension of the former analysis but rather a complementary approach. It would be easy to deal with the dichotomous case for  $\partial=0$  and  $\partial=1$  in the same way. Yet, since these extensions are straightforward, we are content to let it rest with that and to outline the argument merely for one case.

### III.3. Continuous variation of $m$

In the preceding discussion trustworthiness is still characterised in terms of the incentive structure of the original "basic game of trust". Whether or not an individual is regarded as a trustworthy type depends merely on the relationship between  $m$  and  $r-1$ . In the original game of trust this makes good sense since rational behavior in the position of a second

mover is determined independently of the population composition. Rational behavior in the role of the second mover and type completely coincide. For all population compositions  $p$  the trustworthy  $\underline{m}$ -type chooses R while the non-trustworthy  $\bar{m}$ -type chooses E. However, if the game is to be played in the shadow of the courts, it may well happen that type and behavior diverge for some  $p$ .

Though it is true that  $\underline{m}$ -types independently of the population composition will always choose R,  $\bar{m}$ -types, expecting to be sued, should choose R instead of E if  $\bar{m} < \omega(p)$ . Due to  $0 \geq \bar{m}$  and  $C < 0$  relation III.3 implies  $\bar{m} < \omega(1)$ . But  $\bar{m} < \omega(p)$  may also hold good for  $p < 1$ . For such population compositions,  $p$ ,  $\bar{m}$ -types like  $\underline{m}$ -types choose R. For  $\bar{m}$ -types behavior and type can diverge. In the shadow of the courts *behaviorally* monomorphic populations of players at the same time may be non-monomorphic in their type composition.

More generally, if  $p'$  denotes the fraction of players who choose R in the role of the second mover and  $p$ , as before, denotes the fraction of types deemed trustworthy according to the original criterion  $m < r-1$ , then without the shadow of the courts  $p' = p$  would be guaranteed. But if the courts cast their shadow on the original game of trust  $p' \neq p$  may well emerge. In particular,  $p' = 1$  and  $p < 1$  is possible. Then, in the *behavioral* sense of expected play, all players are trustworthy even though  $p < 1$ .

In the shadow of the courts, what is a sufficiently strong conscience  $m$  to affect *behavior* varies with the population composition. This suggests that we interpret trustworthiness in terms of expected *behavior*. It becomes dependent on behavioral expectations and thus on the population composition as characterised by  $p$ . In short, trustworthiness is no longer a *motivational* concept going along with type as such but rather a *behavioral* one based on strategic rationality.

We would not treat judges like ordinary citizens would we not apply the behavioral concept of trustworthiness to the judges as well. But then we must change the view that  $\bar{m}$ -judges are non-trustworthy independently

of their own likely behavior as players and thus independently of the population composition. According to this view  $\bar{m}$ -judges decided on restitution with some *fixed* probability  $q < 1$ . Even if the shirking judges due to  $q \ll 1/2$  were prone to favor the unfair very strongly they did so independently of which behavioral pattern prevailed in the population. Now the "general behavioral pattern" or "the general climate" in society, as characterised by  $p'$  rather than  $p$ , is crucial for determining judges' behavior as well. A frequency dependent motivational factor is superimposed on the factor  $q$  describing the degree to which shirking judges sympathise with unfair players.

Making the distinction between  $p'$  and  $p$  suggests that we restate the condition III.4 (for choosing E) in terms of  $p'$  rather than  $p$ . As is obvious from the derivation of the equivalent relation III.3 the population composition affects relation III.4 only via the judges' decision making. Judges' *decisions* rather than their types matter. Moreover, the function  $\omega(p')$  is monotonically increasing in  $p'$  rather than in  $p$ . The absolute value of the conscience necessary to overcome the basic incentive to exploit a trustful first mover becomes lesser and lesser with increasing  $p'$ . Who is a trustworthy type in terms of her behavior can no longer be decided by simply looking at the relation between  $m$  and  $r-1$  in the original game of trust. The specific value of  $m$  does matter if behavior in the role of the second mover is affected by  $p'$ .

Taking into account the interdependence between  $m$  and  $p'$  it seems appropriate to allow for a mutant space with continuous type variation. Rather than distinguishing once and for all between trustworthy and non-trustworthy types according to the relation between  $m$  and  $r-1$  the line between the "behaviorally" trustworthy and the non-trustworthy should be drawn according to the population composition itself.

To accomplish this, let us assume that the mutant set is the interval  $[\underline{m}, \bar{m}]$  with

$$m \in [\underline{m}, \bar{m}], \quad \underline{m} < r-1 < \bar{m} \leq 0.$$

$\underline{m}$  and  $\bar{m}$ , respectively, are the strongest and the weakest form of a conscience. Allowing for the full range of "intermediate" values for parameter  $m$  we get the game represented in the next figure. The line between  $\underline{m}$  and  $\bar{m}$  indicates that we now take into account all mutants  $m$  with  $\underline{m} \leq m \leq \bar{m}$ .  $G(m)$  is the population distribution based on the density  $f(m)$ . Both functions are common knowledge among the players.

As before the signalling aspects of the game may be neglected. For, whatever inferences the first mover on his second move might draw from observing an act of exploitation this does not affect his expectations. Judges react on behavior rather than on type even if their behavior now depends on how they themselves would have chosen in the role of the second mover and thus on the general social climate.

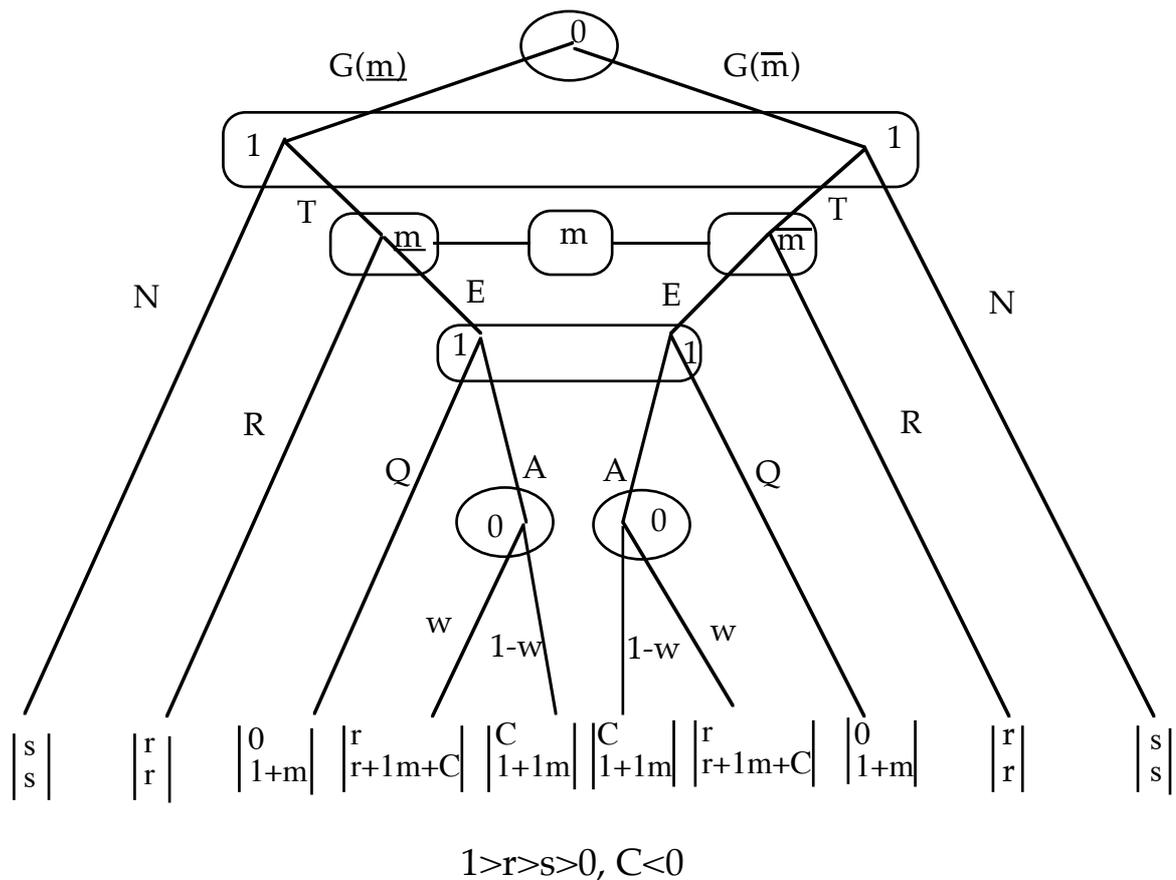


Figure 10

### III.3.1. Solution play

Taking into account our previous discussion of the case  $\partial=0$  and a dichotomous type space the subsequent analysis of the case  $\partial=1$  and a continuous type space can be applied straightforwardly to  $\partial=0$  and a continuum of types. The arguments for conditions of no appeal apply with slight modifications too. Therefore we shall primarily concern ourselves with solution play for  $\partial=1$  under the presumption that the condition for appeal holds good.

If the condition for choosing appeal,  $A$ , applies, then, according to III.4, the second moving player of type  $m$  will choose  $E$  iff  $m > \omega(p)$ . In the dichotomous case either  $\bar{m} > \omega(p)$  did apply or it did not. That is, either all individuals independently of their type were trustworthy or merely the  $\underline{m}$ -types as exogenously defined by  $\underline{m} < r-1$ . However, now the population share of trustworthy individuals is itself endogenously characterised by the behavioral parameter  $p'$ . Who is a (behaviorally) trustworthy type and who is not, only emerges from determining the decisions of  $m$ -types and  $p'$  simultaneously. A player of type  $m \in [\underline{m}, \bar{m}]$  is a (behaviorally) fair type if  $m < \omega(p')$ . But the value of the argument  $p'$  of the function  $\omega(p')$  is itself derived from the type-dependent decisions of all players. How they choose is dependent on their expectations and the latter depend on how they choose.

A conventional rational expectations argument may be used to avoid the circularity in the preceding reasoning from becoming vicious. Assuming self-supporting rational expectations suggests that the players' behavior is guided by parameter values  $m^*$ ,  $p^*$  with

$$m^* = \omega(p^*),$$

the share of  $m$ -types with  $m \in [\underline{m}, m^*)$  is  $p^*$ ,

the share of  $m$ -types with  $m \in (m^*, \bar{m}]$  is  $1-p^*$ ;

$$\text{where } \int_{\underline{m}}^{m^*} f(m) dm = G(m^*) = p^* \quad \text{and} \quad \int_{m^*}^{\bar{m}} f(m) dm = 1 - G(m^*) = 1 - p^*.$$

Technically, this terminology requires that  $m = m^*$  is no atom of the distribution  $G(\cdot)$ . The interval  $[\underline{m}, \bar{m}]$  is then split into two sub-intervals corresponding to classes of individuals who would and of individuals who would not behave fairly. For given  $f(m)$  all types  $m \in [\underline{m}, m^*)$  are fair types who choose R and all types with  $m \in (m^*, \bar{m}]$  are unfair types who choose E in the role of the second mover.

Once the parameters  $m^*, p^*$  are determined the continuous is reduced to the dichotomous case. To the fair and unfair types -- as determined relative to a given  $f$  -- our preceding arguments about solution play

apply. In particular, the condition for appeal becomes  $p^* \in \left( \frac{-q}{1-q} \frac{r-C}{r-C}, 1 \right]$  if  $\frac{q}{1-q} r < -C$ , while players choose Q for all compositions  $p^* < \frac{qr + C(1-q)}{(1-q)(C-r)}$ .

### III.3.2. Evolutionary dynamics and stability

Assume discrete time  $t$  and let  $f_t(\cdot)$  refer to the type-density and  $G_t(\cdot)$  to the type distribution at time  $t$ . The evolutionary dynamics are then characterised by changes of these functions through time. As before, we shall also pay special attention to limiting states of the population composition.

In the preceding discussion, trustworthiness and non-trustworthiness of players did not vary with the population composition. Now it does depend on  $f_t(\cdot)$  how on each round of play of the evolutionary game  $[\underline{m}, \bar{m}]$  is split into subintervals  $[\underline{m}, m^*)$  and  $(m^*, \bar{m}]$  corresponding to trustworthy and non-trustworthy individuals respectively.

Again play in the role of the first mover does not differentiate between types. Differences in different types' success, if there are any, must depend on behavioral differences in the role of the second mover. Behavior in the role of the second mover is determined by the relation between  $m$  and  $\omega(p^*)$ , where

$$p^* = \frac{m^* + (1-r)(1-q) + qC}{(1-q)(1-r-C)}.$$

Considering  $p$ -space under the behavioral interpretation of  $p$  is not too helpful, if meaningful at all. We therefore focus on  $m$ -space considering several distinct generic intervals in which  $m^*$  could be located. Each of the  $m^*$  itself separates two intervals of  $m$ : on the one hand, the interval of the  $m < m^*$  who choose to behave fairly and, on the other hand, the interval of the  $m > m^*$  for whom it is rational to choose unfairly given their conscience parameter.

For constellations for which the condition for appeal is fulfilled the threshold  $\bar{m}$  (to be introduced more formally below) separates two realms of values of  $m^* = \omega(p^*)$ . To the left of  $\bar{m}$  are  $m^*$  for which the unfair fare better. If the interaction starts with such  $m^* < \bar{m}$  then  $m^*$  will tend to decrease even more. That is, a conscience of increasing strength will be necessary to induce fair behavior. If  $m^*$  falls to the right of  $\bar{m}$  initially, then the value  $m^*$  beyond which individuals start to behave unfairly will tend to increase. The realm where the fair are more successful is increasing then as well -- of course, provided that the condition for appeal holds good.

As before, it might facilitate orientation if we pre-view our results graphically.

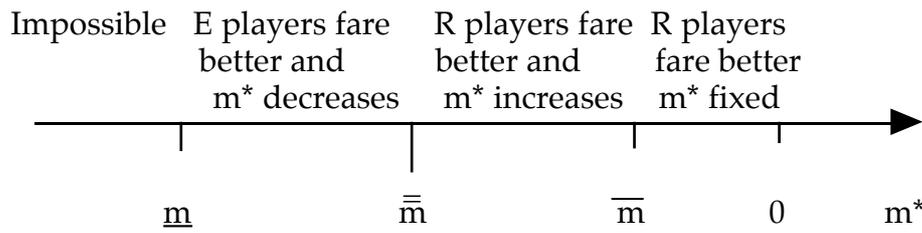


Figure 11

Distinguishing between  $\omega(p^*) \in (\underline{m}, \bar{m})$  and  $\omega(p^*) \notin [\underline{m}, \bar{m}]$ , let us turn now to a more detailed discussion of different "locations" of  $m^*$ . Conceivably the threshold separating the class of those  $m$  that lead to trustworthy from the class of  $m$  that lead to non-trustworthy behavior could in principle be located such that:  $m^* < \underline{m}$ ,  $\underline{m} < m^* < \bar{\bar{m}}$ ,  $\bar{\bar{m}} < m^* < \bar{m}$ ,  $\bar{m} < m^*$ . As a matter of convenience we shall start our discussion with the two locations falling outside the parameter space of  $m$ , i. e. we start with  $m^* \notin [\underline{m}, \bar{m}]$ , and then turn to the central case of  $m^* \in (\underline{m}, \bar{m})$ .

Case  $\omega(p^*) \notin [\underline{m}, \bar{m}]$ :

In this case the threshold separating the behaviorally trustworthy from the non-trustworthy falls outside the realm of possible conscience parameters. For reasons that will become obvious from the subsequent discussion of the two sub-cases of this case, we can neglect the condition of appeal if  $m^*$  falls outside the space of possible values for the conscience parameter  $m$ .

Sub-case  $\omega(p^*) < \underline{m}$ :

$0 > r-1 > \underline{m}$  and  $1 > q > 0$  imply  $(r-1)(1-q) > r-1$ . Since  $-qC \geq 0$  we get  $(r-1)(1-q) - qC > r-1$  and thus, because of  $r-1 > \underline{m}$ ,  $\omega(0) = (r-1)(1-q) - qC > \underline{m}$ . Therefore, due to the strict monotonicity of  $\omega$ ,  $\omega(p) > \underline{m}$  for all  $p$ . The premise  $\omega(p^*) < \underline{m}$  cannot be fulfilled.

Sub-case  $\omega(p^*) > \bar{m}$ :

If  $\omega(p^*) > \bar{m}$  then  $p^*=1$ . For, if  $\omega(p^*)=m^* > \bar{m}$  then all  $m$ -types choose R. The population is behaviorally monomorphic. Since an occasional choice of E would not be systematically related to  $m$  nothing can differentiate between types regardless of whether or not other conditions like that for appeal apply. Moreover, such a population could not be invaded by an  $m$ -type behaving differently unless  $m \notin [\underline{m}, \bar{m}]$ . Since the latter is impossible by assumption any mutant type must also choose R. There cannot be a systematic tendency towards changing the composition of a behaviorally monomorphic population of players who all choose R. Therefore  $p^*=1$  is stable.

In general, we have looked only at generic cases. Since for either  $\omega(p^*)=\underline{m}$  or  $\omega(p^*) = \bar{m}$  we are dealing with the very limits of the mutant space some separate remarks may be appropriate. As far as  $\omega(p^*)=\underline{m}$  is concerned, the argument for the sub-case  $\omega(p^*) < \underline{m}$  would obviously go through for  $\omega(p^*) \leq \underline{m}$  as well. If, on the other hand,  $\omega(p^*) = \bar{m}$  then all  $m$ -types, except the  $\bar{m}$ -types choose R. An  $\bar{m}$ -type would be indifferent between the choice of R and E. Since an occasional choice of E would suffice for some type differentiation to emerge this sub-case is implicitly treated in the discussion of the next case.

Case  $\omega(p^*) \in (\underline{m}, \bar{m})$ :

Since  $\bar{m} > \omega(p^*)=m^* > \underline{m}$  we have  $[\underline{m}, m^*) \neq \emptyset \neq (m^*, \bar{m}]$ . As second movers  $m$ -types with  $m \in [\underline{m}, m^*)$  choose R, while  $m$ -types with  $m \in (m^*, \bar{m}]$  choose E. The population is behaviorally non-monomorphic as far as choice in the role of the second mover is concerned. Therefore, as long as there is a positive probability that first movers either intentionally or by chance choose T this will differentiate  $m$ -types with  $m \in [\underline{m}, m^*)$  from  $m$ -types with  $m \in (m^*, \bar{m}]$  according to their evolutionary success.

For  $m^*$  located within the type-space it matters, whether or not the conditions for appeal are fulfilled. So let us distinguish the sub-cases  $p^* > p^\circ$  and  $p^* < p^\circ$ .

Sub-case  $p^* > p^\circ = \frac{qr + C(1-q)}{(1-q)(C-r)}$  or "appeal":

If the condition for appeal applies then success, as measured in objective payoff, is  $r$  after playing  $R$  and  $w(r+C) + (1-w)1$  after playing  $E$ . An  $m$ -type who is non-trustworthy under a given density  $f_t(\cdot)$  is more successful than a trustworthy one if

$$(III.7) \quad w(r+C) + (1-w)1 > r$$

$$\Leftrightarrow p^*(1-q)(1-r-C) < (1-q)(1-r) + qC \quad \Leftrightarrow$$

$$(III.7') \quad 0 \leq p^* < \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C}$$

$$\text{Define } \bar{m} := \omega\left(\frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C}\right)$$

Note also that the strict monotonicity of  $\omega$  implies

$$m^* < \bar{m} \Leftrightarrow p^* < \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C} \text{ and } m^* > \bar{m} \Leftrightarrow p^* > \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C}$$

These equivalences relating  $m$ -space and  $p$ -space should be kept in mind when interpreting figure 11 above.

In view of III.7' we can distinguish two basic constellations then:

$$(i) p^* > \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C} \quad \text{and} \quad (ii) p^* < \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C}$$

Constellation (i) or "fair fares better":

Observe that  $0 < 1-r < 1$ ,  $C < 0$ ,  $0 \leq q < 1$  imply

$$(III.8) \quad \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C} < 1.$$

The interval  $(\frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C}, 1]$  is non-degenerate. For  $p^*$  from this interval we have  $w(r+C) + (1-w)(1) < r$ . In view of  $\bar{m} \leq 0$ ,  $w(r+C) + (1-w)(1) < r$  implies  $w(r+m+C) + (1-w)(1+m) < r$  for all  $m \in [\underline{m}, \bar{m}]$ . Therefore, obviously, all players irrespective of their  $m$ -type choose to play R if constellation (i) prevails. This, in turn, amounts to  $p^* = 1$ . Therefore there will be no systematic tendency towards changing that behaviorally monomorphic population composition.

Inducing the choice of E instead of R would require a type  $m$  with  $m \notin [\underline{m}, \bar{m}]$ . Therefore, if III.8 holds good  $p^* = 1$  is stable.

Constellation (ii) or "unfair fares better":

As can be seen from III.7' if

$$(III.9) \quad 1-r > -\frac{q}{1-q} C$$

applies then condition III.7 defines a generic interval. Notice also that for sufficiently small  $|C|$  the condition stated as constellation (ii) and the condition for appeal can simultaneously be fulfilled.

Now, if for some point in time  $t$ ,  $p^* \in [0, \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C})$  then  $m$ -types who behave non-trustworthy under the density  $f_t(\cdot)$  are more successful than behaviorally trustworthy types. The density of behaviorally trustworthy will decrease while that related to those who behave in non-trustworthy ways increases. Under density  $f_{t+1}$  the share  $p^*$  of those who behave trustworthy will be smaller than under density  $f_t$ . Moreover, since  $\omega(p^*)$  decreases if  $p^*$  decreases, the value  $m^* = \omega(p^*)$  will be smaller

in  $t+1$  after a decline of  $p^*$  in  $t$ . In  $t+1$  the intrinsic motivation necessary to overcome extrinsic incentives to behave "opportunistically" must be stronger than in  $t$ . The type-interval is split into trustworthy and non-trustworthy types at a point more towards  $\underline{m}$  than before.

In sum, if III.7' and the condition for appeal are fulfilled for  $p^*$  the population share of "behaviorally" trustworthy individuals will tend to decline.

The courts will be involved in the process leading to this decline of  $p^*$  until  $p^* = p^\circ = \frac{qr + C(1-q)}{(1-q)(C-r)}$  is reached. But  $p^* = p^\circ$  is not stable. As is shown in the discussion of the next sub-case, independently of court action  $p^*$  will decline beyond  $p^\circ$  if first movers after being exploited just quit.

Sub-case  $p^* < p^\circ = \frac{qr + C(1-q)}{(1-q)(C-r)}$  or no appeal:

If the condition for appeal is not fulfilled, then the argument outlined in the discussion of constellation (ii) basically applies again. As long as  $T$  is chosen with minimum positive probability the parameter  $p^*$  characterising the population share of those  $m$ -types who act fairly will decline. As shown in the discussion of case  $\omega(p^*) \notin [\underline{m}, \bar{m}]$  above,  $\omega(p^* = 0) > \underline{m}$ . Therefore the value of  $m^* = \omega(p^*)$  declines until a behaviorally monomorphic population composition characterised by a density  $f$  leading to  $p^* = 0$  emerges. For all  $m$ -types  $m \notin [\underline{m}, \omega(p^* = 0)]$  or  $m > m^* = \omega(p^* = 0)$ . Thus, all choose  $E$  and there is no further systematic tendency to change the composition of the population. Therefore a population characterised by a density  $f$  leading to  $p^* = 0$  is evolutionarily stable in the LESS-sense.

In sum, we can characterise evolutionary dynamics and stability for continuous type variation in the following way now:

If the population is behaviorally non-monomorphic initially, that is, if  $p^* \in (0, \frac{1-r}{1-r-C} + \frac{q}{1-q} \frac{C}{1-r-C})$  then the evolutionary process will eventually yield a behaviorally monomorphic stable population composition characterised by a density  $f$  leading to  $p^*=0$  and the use of E by all players.

If the population is behaviorally monomorphic, then, if  $p^*=0$ , initially, a population of players who all are using E and if  $p^*=1$ , initially, a population of players who are all using R are stable.

Again our results, as compared to those for the original evolutionary game of trust, do not suggest that a general crowding out effect exists. Whether or not crowding out in fact can be observed depends critically on the parameter constellation, the initial distribution  $f$ , the value of  $p^*$  implied, and, last but not least, on the bench mark process. As far as this is concerned again attention should focus on the evolutionary game of trust in which some type information of limited reliability is available at some cost. Only if types can ex ante be distinguished to a certain extent, can there be bi-morphic populations with  $p>0$  and thus a crowding out effect at all (cf. again figure 4 above). It is most important to note, however, that in the region where R is universally chosen again the courts can in fact prevent superior evolutionary success of individuals endowed with a relatively weaker -- that is less negative -- conscience parameter  $m$ .

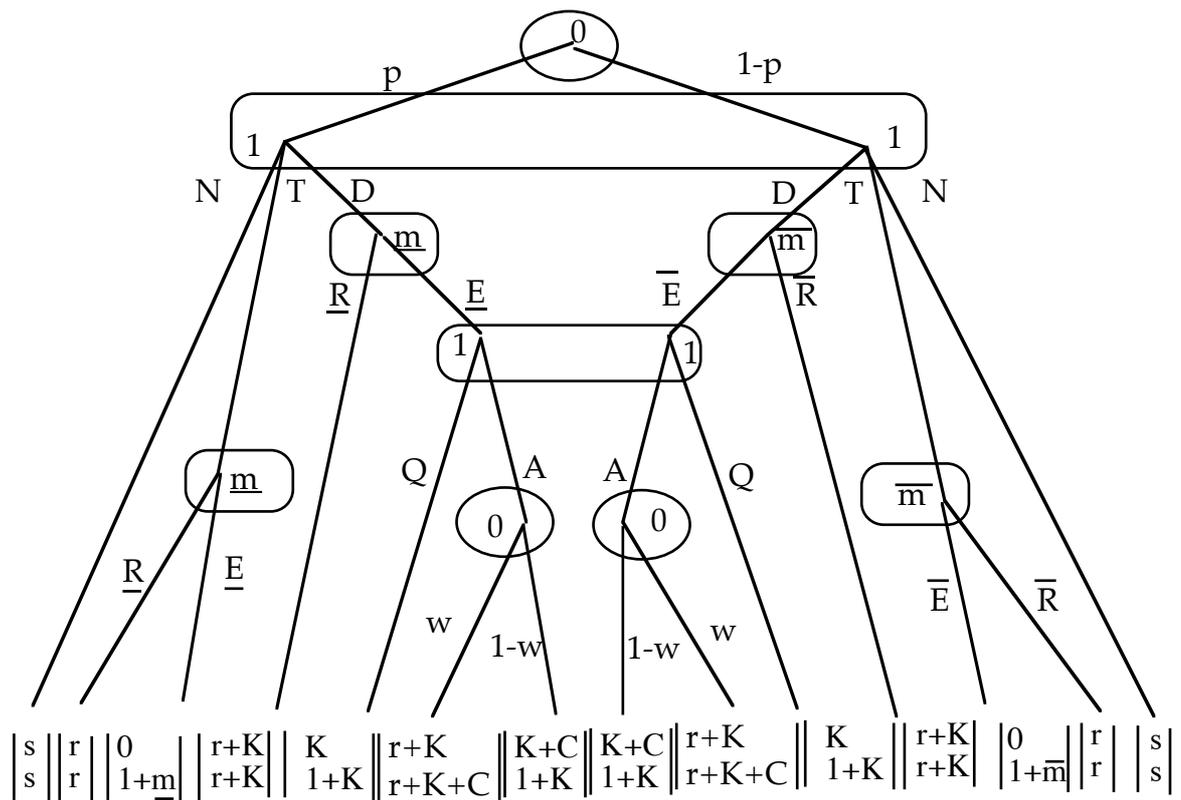
The central result that the introduction of courts can per se improve performance stands up under the behavioral as opposed to the motivational interpretation of trustworthiness. Introduction of a court system can improve overall performance if judges decide according to how they themselves would behave as citizens.

IV. Discussion

In view of formal models like the previously discussed it must be asked how meaningful the abstractions involved are, how robust the conclusions, how plausible the interpretations, and what kinds of broader questions might be posed in the sequel. In our efforts to address at least some of these issues let us start with the question of how robust our findings are if the influence of courts is modelled differently.

IV.1. Alternative modelling of the influence of the court

It is quite plausible that intrinsic motivation is not only suspended after being sued but rather in the very moment that formal contracting procedures are used and thus distrust is shown.



$$1 > r > s > 0, 0 > K, C, \text{ and } \bar{m} + 1 > r > \underline{m} + 1$$

Figure 12

In this game all branches of the tree of the original game of trust show up. The player who is to play as first mover can still choose between N, no trust, and T, trust. Again assuming without loss of generality that player 1 is moving first he has now the additional option of choosing D, or of showing distrust. Distrust is different from N or refusing to interact any further. It relies on formal contracting along with the intention to interact in a more formal setting structured by the threat of extrinsic intervention. A relationship that could have been framed as a (personal) trust relationship by choosing T is framed as a contract or (impersonal) relationship. It seems quite plausible that, after distrust is shown, not only additional costs  $K$  of formal contracting are incurred but also that this suppresses all intrinsic motivational factors typical of informal trust relationships. -- Since the rest of the game tree should be self-explanatory let us directly turn to analysis.

Without determining the solutions of the game under various parameter constellations one can easily see that only the  $\bar{m}$ -monomorphic population composition with  $p=0$  is limit evolutionarily stable. Consider first that player 1 after being exploited would rationally choose Q. In that case the choice of D is dominated by T for all  $K < 0$ . Thus the old game of figure 3 emerges. In this game, according to our previous results, only the  $\bar{m}$ -monomorphic population composition with  $p=0$  is limit evolutionarily stable. This is not changed by the fact that with a small probability, D is chosen even if non-D is the rational decision.

If D and A are chosen basically the game of figure 5 for  $\partial=0$  emerges -- though now  $m$  is completely eliminated. With respect to strategic choices,  $K$  is merely an additive constant that can be neglected too. As before, different types are not differentially affected as far as their "objective" payoffs are concerned. It seems that all population compositions for which the first moving player appeals if he comes to move again could be weakly stable at least in principle in this case. Note however, that first movers might make a mistake T once in a while. In that case the unfair type of the second mover would fare better while along all other

branches of the tree, even if slight mistakes occur, both types would do equally well. Therefore in this tree we have only  $p=0$  as a limit evolutionarily stable population composition.

It should be noted, however, that in the realm of population compositions  $p > \bar{p}(C')$  the courts can affect the evolutionary dynamics. In the original game of trust with costly type information of imperfect reliability everybody would simply show trust in the role of the first mover. For, then the non-trustworthy are too few to render incurring the costs of discrimination worthwhile. Thus, if  $p > \bar{p}(C')$  the non-trustworthy are more successful (cf. again figure 4). In that case, for  $|K| < |C'|$ , usage of the court system can at least conceivably slow down the elimination of the trustworthy. Instead of getting always a chance to exploit in the position of the second mover the non-trustworthy have an advantage only after occasional choices of T as long as the condition of appeal applies and the courts are used after formal contracting.

Still, the occasional advantage of the trustworthy cannot be eliminated. Nevertheless, the trustworthy are not crowded out *by* relying on the courts but are rather eliminated through interaction outside the shadow of the courts. If the framing effect exerted by the court is as described in figure 12 courts are neutral with respect to crowding out. On the other hand, if framing works the way assumed in figure 12 then the courts cannot be instrumental to reducing evolutionary pressure against the trustworthy. The courts are neutral in this respect, too. From this it is obvious that we must hedge a bit our previous claim that the introduction of the courts per se can work to the advantage of the more trustworthy. What we can say, though, is the following: Introduction of the courts can improve overall performance even if judges are no better than the rest of society provided that players frame the situation in a suitable way.

## IV.2. Reliance on extrinsic enforcement as a general strategy

It may well be that in the game of trust under the shadow of the courts overt behavior better complies with standards of fair behavior than in the game of trust where merely some type information at some cost is available. Behaviorally monomorphic populations of individuals who all play R in the role of the second mover may emerge in case  $\partial=1$ . This holds good also if there is a continuum of m-types who decide on the basis of a behavioral standard of trustworthiness. In social interactions in which trust relationships matter relying on the courts may be preferable to a purely moral guidance and control. Even if framing effects work along lines as sketched in figure 12, introducing a court system at least will not harm the more trustworthy. Even if judges are not more trustworthy than the population at large having them can improve overall trustworthiness.

Beyond that, the models suggest that crowding out through formal enforcement is a real danger only if there is quite a bit of both, fairness and unfairness. In such mixed situations in which most people quite naturally might think that formal enforcement may be *the* way to stop a further "decline of morals" the crowding out effect may in fact be a real danger. If in such situations the cost parameter is "wrongly" set and people rely on the courts rather than on their own information technologies this may indeed destabilise a former stable mixed equilibrium  $p=\bar{p}(C')>0$  as characterised in figure 4 above. If individuals do not switch back to their information technologies such a destabilisation may lead to a complete "break down of morals". In fact, as our analysis of the game of trust with some type information has shown, switching back may be impossible after  $p$  has declined beyond the threshold of  $\underline{p}(C')$ .

The argument must, of course, also be seen in light of the fact that interactions are in general embedded in a larger context (cf. on embeddedness Granovetter 1985). There may be spill over effects on other interactions. These effects will be relevant whenever the strength of the conscience in one class of interactions is related to the strength of the

conscience for another kind of interaction. As psychology tells us, similarity of interactions is important. Closeness in this sense may be relevant for the strength of expected spill over effects. If individuals cannot arbitrarily shift their behavioral gears, if they cannot completely separate their behavior in different roles then such effects will be very likely.

The factors mentioned in the preceding paragraph may lend much greater importance to crowding out than it might otherwise have. It should be noted again, however, that the argument from spill over effects cuts both ways. As our models clearly show, the courts may very well stabilise a population composition that otherwise would suffer from a crowding out effect. In particular, they can eliminate negative spill-over effects on other interactions deemed similar by ordinary players.

The latter somewhat speculative remark seems to coincide well with folk psychology. For instance, if individuals make the experience that they can get away with almost any traffic violation this is expected to have some influence on the general climate of law obedience. It may be necessary to enforce the law more severely in order of not inducing the view that the law does not matter in other realms as well. Closer to our main concerns here, those who have experienced again and again that they cannot trust promises may change their own views on how binding formal contracts are. Here stronger formal enforcement may be the key to preventing the elimination of trustworthiness.

Many people are tempted to accept the preceding line of argument throughout. But this goes too far as well. As demonstrated by our models under certain parameter constellations relying on courts after interacting rather than relying on inspection and type information before interacting may have a negative crowding out effect on moral dispositions. No sweeping general claims are warranted. Studying models like the preceding is a way to enhance our awareness of what *can* happen but it may also suggest some further thoughts.

#### IV. 3 Crowding out and subsidiarity

Crowding out may be seen in a much broader perspective than indicated so far. It is closely related to the ongoing debate about a more communitarian and a more liberal view of social organisation. With respect to the preceding models this is borne out most clearly if we focus on two quite distinct notions of subsidiarity. The locus classicus for the principle of subsidiaric organisation of society is the papal encyclical *quadragesimo anno*. This encyclical states in article 79 (cf. for instance the edition Carlen 1990):

"Still, that most weighty principle which cannot be set aside or changed, remains fixed and unshaken in social philosophy: just as it is gravely wrong to take from individuals what they can accomplish by their own initiative and industry and give it to the community, so also it is an injustice and at the same time a grave evil and disturbance of right order to assign to a greater and higher association what lesser and subordinate organizations can do."

#### Or with Abraham Lincoln's words

"The legitimate object of government is to do for a community of people whatever they need to have done but cannot do at all, or cannot do so well for themselves in their separate and individual capacities. In all that people can do individually as well for themselves, government ought not to interfere."

Of course, normally private contracting would be regarded as form of doing things for oneself without state intervention. It is regarded as fully compatible with the subsidiarity principle then. However, as far as the courts are seen as part of the state it is not true that people fully organise things themselves if they do so through formally enforced private contracting. (Noting this is also of considerable importance with respect of interpreting the role of the subsidiarity notion in the Maastricht treaty.)

The ability of individuals to do things for themselves without formal enforcement procedures and without even private contracting may be viewed as subsidiarity in the narrow sense. We suppose that the catholic church had this communitarian kind of subsidiarity in mind in qudaragesimo anno. If that is so, then introducing court enforcement per se shall crowd out other forms of social organisation. But it will not necessarily eliminate potential advantages of trustworthy individuals. More generally, the preceding line of argument suggests that we should very carefully distinguish between crowding out of intrinsic *motivations* or of trustworthiness and crowding out of such things as certain *forms of organising social interaction*. The latter might not affect the underlying moral dispositions at all -- at least not necessarily.

Lincoln's view is presumably closer to the classical liberal concept of (subsidiaric) organisation of society. According to this view formal enforcement through courts is compatible with subsidiaric organisation as long as it is derived from private contracting. Moreover, it is regarded as compatible with maintaining the moral motivations or dispositions of the citizens. Clearly the German ordo liberals endorsed this view. Even though some of them had strong catholic leanings at the same time, they were in favor of what Franz Böhm called the "Privatrechtsgesellschaft" (the private contract society) as a *moral* ideal. As our analyses show they could well do that without having to fear that introducing private contracting with "formal" court enforcement would per se be detrimental to "morals" in general. It might however eliminate non-contractual forms of organisation and in that sense lead to that kind of "coldness" that so-called capitalist forms of organisation are often blamed of.

## References

- Carlen, C., Ed. (1990). *The Papal Encyclicals*. Pierian Press.
- Frank, R. (1988). *The Passions within Reason: Prisoner's Dilemmas and the Strategic Role of the Emotions*. New York, W. W. Norton.
- Frey, B. S. (1997). "A Constitution for Knaves Crowds Out Civic Virtues," *The Economic Journal* forthcoming(July)
- Granovetter, M. (1985). "Economic action and social structure: The problem of embeddedness," *American Journal of Sociology* 91(3): 481-510.
- Güth, W. and H. Kliemt (1994). "Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes," *Metroeconomica* 45(2): 155-187.
- Güth, W. and H. Kliemt (1995). "Evolutionarily Stable Co-operative Commitments," *Humboldt University Discussion Paper- Economics Series* 53
- Kahneman, D. and A. Tversky (1984). "Choices, Values and Frames," *American Psychologist* 39(April): 341-350.
- Lindenberg, S. (1983). "Utility and Morality," *Kyklos* 36/3: 450 ff.
- Robertson, D. H. (1956). *Economic Commentaries*. London, Staples Press.
- Selten, R. (1988). "Evolutionary Stability in Extensive Two-person Games -- Corrections and Further Development," *Mathematical Social Sciences* 16(3): 223-266.

The paper

Trust in the shadow of the courts if judges are no better still contains complete trivial derivations of formulas. The trivial parts shall be eliminated eventually but are left in for the convenience of checking on them more easily. After that the final formatting will be done.