Tilburg University

Trust in the Shadow of the Courts

Brennan, G.; Güth, W.; Kliemt, H.

Publication date:
1997

Citation for published version (APA):
Brennan, G., Güth, W., & Kliemt, H. (1997). *Trust in the Shadow of the Courts*. (CentER Discussion Paper; Vol. 1997-89). CentER, Center for Economic Research.

Trust in the shadow of the courts[+]

by

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

Research School of Social Sciences, Australian National University, Canberra*; Economics Department, Humboldt University, Berlin**; Philosophy Department, Gerhard Mercator University, Duisburg***

Abstract

If contract enforcers must be randomly selected from the same population and thus are as opportunistic as ordinary traders could a system of adjudication nevertheless increase the degree to which contractual obligations on large anonymous markets are fulfilled? Adopting an indirect evolutionary approach with endogenous preference formation it can be shown that without superior behaviour of adjudicators an adjudication system can induce untrustworthy traders to behave as if trustworthy. However, in the presence of occasional mistakes adjudication will merely slow down but not fully eliminate the evolutionary advantage of untrustworthy types. Only if arbitrators become judges who receive a fixed income occasional mistakes will not favour untrustworthy types. But even then under non-optimal court politics and unfavourable parameter constellations in a low trust environment the introduction of courts may in fact contribute to the crowding out of the trustworthy.

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

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

[+] The authors would like to express their gratitude to the CentER for Economic Research, Tilburg University, (Güth and Kliemt) and All Souls College, Oxford, (Brennan) for their hospitality as well as for their financial and intellectual support during the period in which the basic work on the paper was done. Particularly helpful was a referee report on the first version of the paper that we received after submitting it to the working paper series of the CentER of Economic Research.

I. Introduction

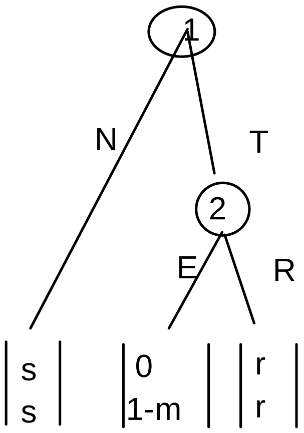
The traditional economic model of competitive markets in which large numbers of anonymous traders engage in mutually beneficial one-off transactions under a legal umbrella of perfectly specified contracts assumes away all problems of trust. In fact, however, such problems are endemic. Even in bilateral on-the-spot exchanges of goods of commonly known quality it is not possible for both parties to make the execution of their own contractual promises contingent on the other party's prior performance. Logically at least one of the partners must be induced to take a risk and to fulfil his part without knowing whether the other is doing his. This is the basic 'trust predicament' that lurks in the background of all transactions between opportunistically rational individuals.

In this paper, we offer an account of how, in the face of the trust predicament, large scale markets can serve a useful function and can be maintained among rational actors. We use an indirect evolutionary approach to show that institutions of enforceable adjudication in themselves may enable higher levels of contract compliance than would obtain in their absence even though adjudicators are no better behaved than ordinary traders. In section II we introduce our basic methodology, lay out the trust predicament and briefly sketch some previous results concerning evolutionarily stable equilibria in the absence of adjudicative institutions. Section III introduces our model of court behaviour, and isolates the values of parameters under which the courts can have behavioural effects. The impact of these behavioural effects on the evolutionary dynamics and stability of the population composition are discussed in section IV. Section V offers some broader conclusions.

II. The indirect evolutionary approach to trust

Within standard rational choice analysis, preferences/utility functions are exogenous: preferences may conceivably change, but not in a way that is interior to the models. The indirect evolutionary approach as conceived here (on this originally Güth and Yaari 1992) by contrast treats utility functions as (partly) endogenous. Utility functions are subject to an evolutionary process insofar as the type composition of a population of bearers of different utility functions evolves through time. Analysing this process we can go some way towards explaining the emergence of preferences and thus move beyond the limits of conventional rational choice analyses. In this paper, we shall be concerned with an application of this technique to 'the problem of trust'.

The basic *trust game* illustrated in Figure 1 represents the social predicament with which trust is expected to deal.



1>r>s>0, 1/2<r

Figure 1

The interaction is well-known and can be described briefly. There are two players 1 and 2. Player 1 chooses first and may either 'trust' 2 (choose T) or 'not trust' 2 (choose N). In the latter case both players receive s (>0) and the game ends. In the former case player 2 gets to choose between 'exploit' (E) and 'reward' (R). If R is chosen, both players receive r (1>r>s). If E is chosen, player 1 receives 0, but player 2 receives a pay-off of 1-m, where 1>m>0. Assuming that the game tree and thus the values of m, r, s, 1, 0 are common knowledge among the players the equilibrium outcome of the game is (T; R) if m>1-r, and N if m<1-r. In the former case the pay-off vector is (r, r) and Pareto efficient while in the latter case the pay-off vector is (s, s) and thus Pareto dominated since (r, r)>(s, s) by construction. The Pareto efficient outcome is made inaccessible by player 2's rationality.

We shall think of the parameters r, s, 1, 0 as based on some 'objective' aspects of the real world, like resources directly related to evolutionary success. The parameter m is different in this respect: it is a purely 'subjective' motivational factor that does not represent an objective aspect of the real world but rather an intrinsic evaluation of the E strategy. For convenience we shall refer to it as the 'conscience parameter'. In other contexts it is useful to let m range over a non-empty interval of parameter values but for the purposes of the exercise here it is sufficient and simplifies considerably to let m take only two possible values: $\underline{m}$ (>1-r>0) and 0. These values correspond to two player types: a trustworthy type (m=$\underline{m}$); and a non-trustworthy type (m=0).

The evolutionary mode of analysis we adopt involves conceptualising social interactions as an evolutionary process. On each round of the evolutionary process players are independently 'drawn' from an appropriately large population in which there is a fraction p [0, 1] of trustworthy $\underline{m}$-types, and randomly matched to play the basic trust game. The rules of the game and the population composition parameter, p, are common knowledge among the players. Before matching, the players do not know whether they are going to play in the role of the first- or second-mover. They are assigned their roles as first- and second-movers, respectively, with equal probability.

In the role of the first-mover, player type is irrelevant. There is no independent disposition to trust: first-movers independently of their own type trust solely on the basis of their best assessment of *second-mover* type. Only in the role of the second-mover is behaviour type dependent and differential evolutionary success of different types depends solely on pay-off differences in that role. In this connection we can distinguish two fundamentally different 'polar' cases: that where the first mover knows the value of m for the second-mover; and that where the value of the second-mover's m is unknown to the first-mover. In the first 'complete type information case', it is clear that being a non-trustworthy type is evolutionarily disadvantageous. Trustworthy types in second-mover roles are trusted and receive a pay-off of r; non-trustworthy types in second-mover roles are not trusted and receive a pay-off of s (<r). Since both types do equally well in first-mover roles, the equilibrium value of p in the evolutionary setting is p=1. Moreover, trustworthiness is a strictly dominant strategy in the evolutionary game; and the 'universal trustworthiness' equilibrium is evolutionarily stable in the

very strong sense that groups of non-trustworthy types, even if very large, cannot invade a population of trustworthy types.

In the opposite polar 'private information' case, there is no information as to second-mover type, beyond knowledge of p. Clearly, if p is sufficiently high initially, first-movers will rationally choose to trust. Hence, non-trustworthy second movers will receive 1 on all rounds of play in which they are assigned the role of the second-mover, while trustworthy second movers will then receive r. Since r<1, non-trustworthy types will do better than trustworthy ones. Once p falls below a threshold level (p= **Error!**), first-movers will not rationally trust. The non-trustworthy types will fare no better as second-movers in the basic game, than do trustworthy types. In this sense, it might seem that the equilibrium value of p from above is **Error!**. However, if we (plausibly) allow for occasional lapses by first-movers, we should apply the concept of a 'limit evolutionarily stable strategy' (LESS - see Selten 1988). If, in the range 0• p<**Error!**, first-movers occasionally fail to choose N then the untrustworthy second-mover's pay-off is 1 while a trustworthy second-mover will only receive r (<1). Thus as long as mistakes cannot be ruled out, the non-trustworthy will have an advantage over the trustworthy types and p will eventually be driven to zero. Accordingly, in the 'private information' case, over the range p>**Error!**, there is 'strong' or 'strategy driven' convergence to p=**Error!**, and for p<**Error!**, 'weak' or 'mistake driven' convergence to p=0.

The precedingly sketched analysis of the polar cases is straightforward and simple. But, clearly, the more interesting cases lie in the range between the 'complete type information' and 'private type information' extremes -- what we shall call the 'partial (type) information cases'. These can be modelled in a variety of ways. One particularly instructive approach involves a 'technology' that provides to the first-mover specific information of reliability μ (1/2• μ• 1) about the type of the second-mover with whom he is matched. The technology's type signal is available at cost C (• 0). The parameter μ is the probability that the signal is correct. Together the two parameters μ and C determine whether or not it is worthwhile for rational first-movers to make use of the technology and thus to acquire specific information about the particular second-mover's type.

We think of a technology in the widest sense of that term here, e.g. the possibility of using an inquiry agency, of keeping track of other individuals' reputations, etc. This technology influences the evolutionary process in the intermediate, partial type information case. We here describe the evolutionary process somewhat further to set a benchmark against which the subsequent discussion can be interpreted (for a fuller account and analytical details see Güth and Kliemt 1995).

For all values of the population composition parameter p the two characteristic parameters µ and C of the 'C, µ'-technology determine whether or not it is worthwhile for rational first-movers to acquire specific information about the second-mover's type. Initially, we take the parameter µ as given. The effect of the availability of the technology on the evolutionary dynamics and limit evolutionarily stable population compositions for different initial values of p can then be depicted graphically as in Figure 2 (for an alternative intuitive presentation of the same basic idea cf. Frank 1988)

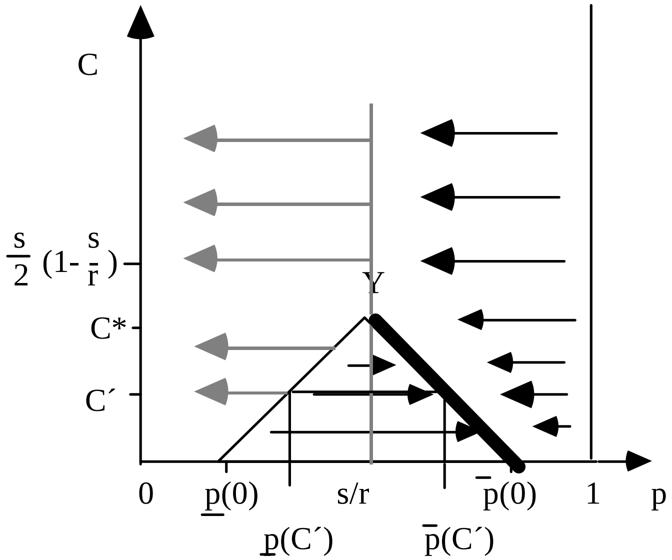


Figure 2

Consider, the case where the technology costs C'. For initial p, there are three ranges of interest: $p< \underline{p}(C')$, p;¯ (C´) <p, $\underline{p}(C')<p<$p;¯ (C´) .

p<$\underline{p}$(C'). No first-mover uses the technology. The proportion of trustworthy persons is too small for the number of trustworthy types identified by the technology to be large enough to justify the cost, C'. Without specific type information no one in the role of the first-mover rationally trusts (since p < **Error!**). Accordingly, trustworthy and non-trustworthy types in second-mover roles fare equally well under rational play in the basic game. However, if first-movers make mistakes and do trust occasionally, untrustworthy types do better than trustworthy ones in the role of the second-mover: hence, the limit evolutionarily stable equilibrium value of p is zero. The dotted directional line (arrow pointing left) at cost level C' indicates the weak (i. e. mistake-driven) convergence of p to zero.

p>p;¯(C') . No first-mover uses the technology. The proportion of trustworthy persons is sufficiently large that first-movers are better off avoiding cost C' and 'trusting to luck'. Without specific type information everyone in the role of the first-mover rationally trusts (since p > **Error!**). Untrustworthy types do better than trustworthy ones in the role of the second-mover: hence p decreases. The undotted directional line (arrow pointing left) at cost level C' indicates the fast (i. e. strategy driven) decline of p to p;¯(C') .

$\underline{p}$(C')<p<p;¯(C') . Every first-mover uses the technology. First-movers trust if and only if the technology indicates that the second-mover is of the trustworthy type. Since for $\underline{p}$(C')<p< **Error!**no first-mover rationally trusts without specific type information, in this range there is more trust shown than in the absence of the technology. Obversely, over the range **Error!**<p<**Error!**, there is less trust in the presence of the technology than there would be in its absence. Over the entire interval, ($\underline{p}$(C'), p;¯(C')), trustworthy types in the second-mover role do better than non-trustworthy ones: hence p increases to p;¯(C') and the convergence is strong as indicated by the undotted directional line (arrow pointing right).

Consider now the evolutionary stability of equilibrium values of p. For C=C', initial values p [0, ($\underline{p}$(C')) are attracted to the limit evolutionarily stable equilibrium p*=0 while initial values p ($\underline{p}$(C'), 1], are attracted to the evolutionarily stable equilibrium p*=p;¯(C) . In Figure 2 the equilibrium values p*=p;¯(C) of p as C changes are given by the heavy line from p;¯(0) to Y. This line is the locus of all (C, p^*) combinations with a positive dynamically stable value p^* of p. For given μ the combination $Y=(C^*,$ **Error!**) indicates the maximum cost C^* for which the 'C, μ'-technology may conceivably be used.

The preceding discussion (and Figure 2) was based on a particular value of the reliability parameter μ. If μ is increased, the locus of possible (C, p^*) equilibria is a line to the right of the heavy line in Figure 2. In the limit, as p approaches unity, the relevant locus is the straight line running from $(C=0, p=1)$ to ($C=$**Error!**, $p=$ **Error!**) while the line ((0, 0), (**Error!**, **Error!**)) forms the corresponding lower boundary of the attractor set.

The basic lesson to be derived from the model is that, if there exists a technology for acquiring specific information about second-mover type which is sufficiently accurate and not too expensive, an evolutionarily stable equilibrium with a positive proportion p (• **Error!**) of trustworthy persons can emerge. However, except in the limiting case of costless, perfectly reliable specific type information no such equilibrium will be characterised by universally trustworthy behaviour or universally trustworthy persons. The plausible, intermediate cases are such that there are always some 'good' and some 'bad' guys around.

The case in which there is a (costly) technology that reveals specific type information ex ante or before the basic game is played involves, of course, already some departure from the idealisation of anonymous markets. Is this inevitable for markets to work? Is it feasible to replace that technology by more formal controls that are based on ex post information about behaviour and thus are compatible with the assumption of anonymity *before* trade? If we rule out arguably implausible assumptions about the motivations and behaviour of adjudicators -- and specifically if we reject the presumption that adjudicators (or 'judges') are 'better than the rest of us' -- could institutions of adjudication and enforcement (i.e. prototypical courts) still increase the extent of market trade? These are questions we engage in what follows.

III. Courts, Enforcement, Trust and the Basic Interaction

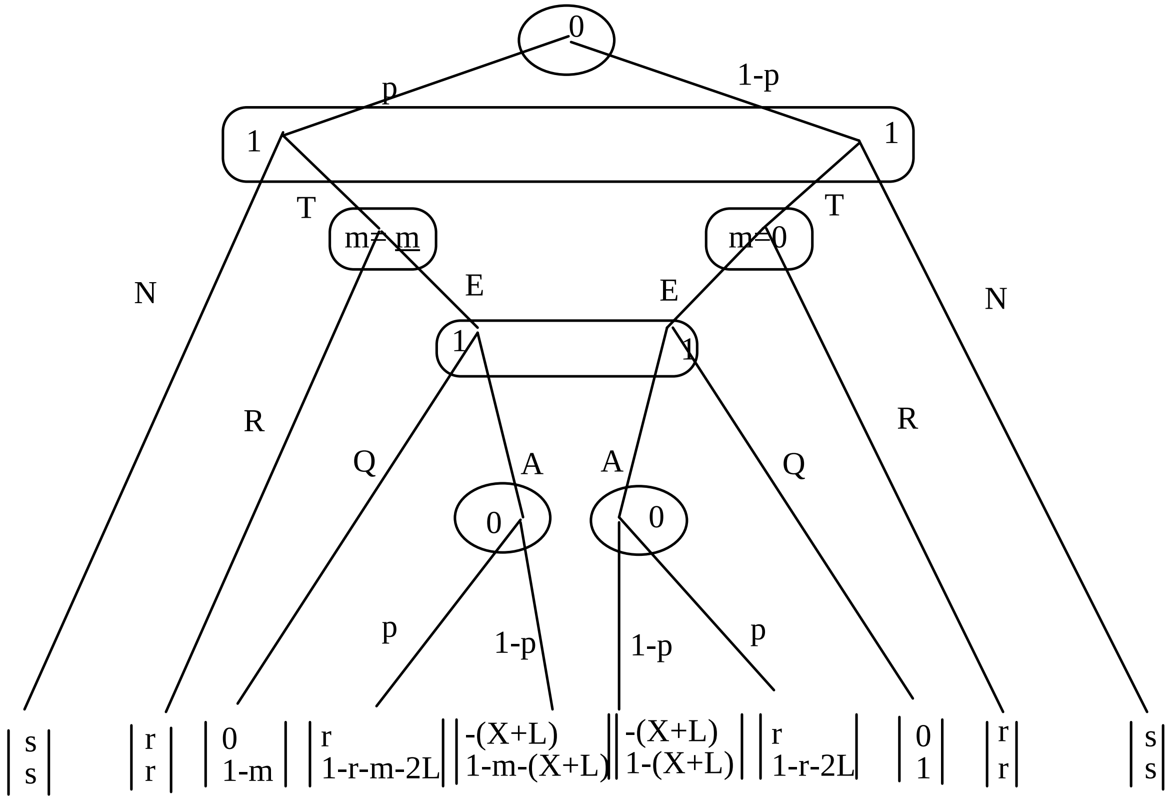
With the foregoing model and considerations as background, we now seek to study the effect of introducing institutions of adjudication and enforcement on how the basic game of trust is played. The formal enforcement of institutional rules may affect not only market behaviour but also the composition of the population of market participants including the enforcers. Our attention will be directed at both behavioural and motivational aspects. We shall be interested in two general questions. First, what are the *behavioural* effects of the court structure, under various values of p, given that adjudicators are drawn randomly from the same population as the players? Second, in the light of these behavioural effects, what values of various parameters, if any, are consistent with which *evolutionarily stable values of the population composition parameter p* ?

A couple of preliminary observations will help limit the terms of the discussion. In particular, it should be clear that the assumptions surrounding the behaviour of the courts -- what they can and cannot do -- are crucial. We shall make three specific assumptions here. First, the operation of the courts will be taken to be *reactive* in the sense that courts intervene only if called into play by one of parties to the basic game of trust. Second, trustworthy persons in their role as adjudicators always find in favour of the 'exploited' party (if there is one). They fix the cost of litigation, 2L (>0), on the 'exploiting party' if there is one, and equally, L, on each party if there is no 'exploiting party'. Third, untrustworthy types as adjudicators decide the case arbitrarily and impose, beyond L, an additional cost on both parties to the dispute of which only the expected amount, 2X, is known ex ante. We shall take it that the *expected* value for each player under an untrustworthy adjudicator is the pay-off in the substantive game minus (X+L), where X (• 0) is that player's expected share of the 'exploitation' that a rational egoistic adjudicator exacts.

This set of assumptions allows us immediately to make one important simplification -- namely, only 'trusting' first-movers will ever rationally appeal, and they will only appeal if they met a second-mover who behaved 'exploitatively'. A brief rehearsal of the various cases is sufficient to establish this proposition. First, if the first-mover chooses N or if the second-mover proves trustworthy, neither party will have an incentive to appeal to the courts. There are net expected costs of doing so: a cost of L to each if the adjudicator is trustworthy; and an expected cost of (X+L) to each if the adjudicator is untrustworthy. Thus, only if the first-mover trusts and the second-mover exploits can there be any possibility of the courts playing a role.

This is the force of the assumption of 'reactive' procedures of adjudication. It greatly simplifies the analysis and at the same time eliminates active or what may be called 'Leviathan' courts which infringe on the property rights of subjects without restraint. This assumption seems legitimate for an analysis whose focus is on contracting among partners when basic property rights are in place. To put the point slightly differently, of the three elements of Humean 'natural law', "the stability of possession, its transference by consent, and ... the performance of promises" (Hume 1739/1978, *treatise*, *b*ook III, part ii, sect. VI), the institutions of adjudication are conceived here to take the first two elements as given and deal only with the third one. However, this construction is taken not to rule out some appropriation of resources by 'untrustworthy' adjudicators when they are activated by appeal.

On this basis we can depict the essential features of the total interaction in figure 3.



1>r>s>0, r>1/2, m>0, L>0, X• 0, 1• p• 0

Figure 3

The game commences with nature's choice of second-mover type, who is 'trustworthy' with probability p. Then at the first stage of the game the first-mover, knowing p but ignorant of whether m=m or m=0, chooses between N and T. After N both players receive s. After T, the second-mover chooses between R and E. If R is chosen by the second-mover both players receive r. So far the game, with associated pay-offs is as for the basic trust game. However, after E the basic game is modified. The first-mover has a further option: if the second-mover has chosen E, the first-mover can either be quiescent, Q, or appeal to the court, A. After Q the pay-offs are the same as in Figure 1 after the play of T and E. After A nature chooses an adjudicator who will be of trustworthy type with probability p, and of untrustworthy type, with probability (1-p). There are two possibilities: First, the adjudicator is trustworthy and the first-mover has his promised reward (r) restored. Then the trustworthy type of the second-mover receives (1-m-r-2L) while the untrustworthy receives (1-r-2L). Second, the adjudicator is untrustworthy in which case the first-mover's expected return is -(X+L). The trustworthy type of the second-mover then receives (1-m-(X+L)) and the untrustworthy (1-(X+L)).

Now, by assumption, for trustworthy players

$$r> (1-m)$$

therefore, since, also by assumption, (r+2L)>0 and (X+L)>0

$$r>(1-m)-(r+2L) \text{ and } r>(1-m)-(X+L).$$

So a trustworthy type in the role of the second mover always chooses R and our attention can focus on the case of an *un*trustworthy type in the role of the second-mover as depicted on the right-hand side of Figure 3.

We can focus initially on the issue of whether an exploited first-mover will appeal to the courts or not. Clearly, if it is not rational for the first-mover to appeal (i.e. to choose A over Q) then the courts cannot exercise any influence on the game at all: the interaction reverts to the basic trust game with private type information and the previous discussion of evolutionary stability in that extreme case tells all. Accordingly, a critical parameter in the system is the value of p such that it pays player 1 to choose A over Q (after moves T, E). This value of p is such that:

$$pr - (1-p)(X+L) > 0$$

$$\text{or } p > \textbf{Error!} \qquad (1).$$

Denote the value of p for which (1) becomes an equality as

$$p^{;A} := \textbf{Error!} \qquad (2).$$

If $p < p^{;A}$, exploited first-movers do not appeal and the basic interaction is strategically equivalent to the basic game of trust without institutions of adjudication previously discussed in section II. Therefore from that discussion we can directly infer that trustworthy types will eventually be driven out if $p < p^{;A}$. However, if $p > p^{;A}$, exploited first-movers appeal. In this case the appeal possibility changes the incentives for untrustworthy second-movers and the system of adjudication imagined here can conceivably have an impact on rational play and consequently on the population composition.

Note, at the outset, that any appeal to the courts is bound to be costly to exploiters: either they get a trustworthy type as adjudicator in which case they will lose an amount r to compensate the exploited party, plus the full costs of 2L; or they will get an adjudicator of the untrustworthy type, in which case they can expect to retain their exploitative pay-off, but lose their share L of the costs of the trial plus the expected rent, X, to the untrustworthy adjudicator. (This reasoning also confirms the modelling assumption that second movers will never appeal to the courts and explains why no corresponding moves show up in the game tree.) Untrustworthy second-movers can, however, avoid the expected cost of court action by fulfilling the terms of the contract in the first place, in which event they receive a pay-off of r. In short, assuming that (1) obtains, it will pay an untrustworthy second-mover to exploit only if:

$$(1-p)\ (1-(X+L)) + p(1-r-2L) > r \ (3)$$

Now, by assumption (see Figures 1, 3)

$$(1-r)< r \qquad (4)$$

so, in particular, 1-r-2L < r. Consequently, for (3) to hold it is necessary that:

$$(1-(X+L)) > r \qquad (5)$$

$$\text{or } (1-r) > (X+L) \quad (6).$$

Inequality (6) -- and hence (3) -- is quite a stringent condition. It can only be satisfied for low values of X and L. Now, (3) can be rewritten as:

$$\textbf{Error!}> p \ \ (7)$$

which can be used to define a threshold of p -- called $p^{;R}$ -- such that for all p greater than $p^{;R}$, an untrustworthy second-mover would rather comply and choose R than to face the courts. Accordingly,

$$p^{;R} := \textbf{Error!} \quad (8).$$

It may be helpful to depict these two conditions in terms of the relation between X and p for given values of L and r. This we do in Figure 4 (a, b, c). The values of L and r for which the relations are derived are indicated on the diagram in each case.

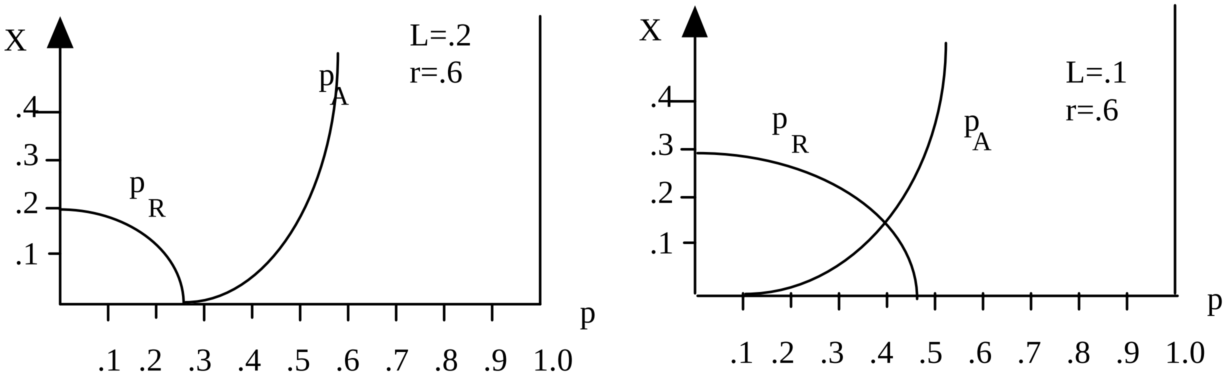


Figure 4a

Figure 4b

Figure 4 (a, b)

The line p ;A divides the plane into two portions: to the right of p ;A , exploited second-movers would appeal; to the left, the courts would not be brought into play. The line, p ;R , also divides the plane into two portions: to the left of p ;R are combinations of X and p such that, if an appeal is expected, rational untrustworthy second movers will still exploit; to the right of p ;R , rational untrustworthy second-movers would rather comply than face the courts. Note that for the parameter values in Figure 4 a), there are no values for which p > p ;A does not ensure p > p ;R ; that is, since in these cases first-movers would appeal second-movers would not exploit them in the first place. The implications of all other possible locations of p can be analysed in a straightforward way as well. However, we cannot rule out the possibility that p ;R > p > p ;A (see Figure 4 b) . In this more complicated case, the proportion of trustworthy persons is such that the first-mover will rationally appeal if exploited, and this fact does not eliminate the incentive for untrustworthy second-movers to exploit.

Now, for this to affect the evolutionary process more frequently than in those instances brought about by occasional mistakes of the first mover, a further condition must be met. Knowing that the untrustworthy type in the second-mover role will rationally exploit, notwithstanding the fact of appeal, will a first-mover rationally trust? She will if her expected pay-off from trusting exceeds s. Her pay-off from T is:

$$p\ r + (1-p)\ [pr - (1-p)\ (X+L)] \qquad (9)$$

and the condition under which the first-mover will trust is:

$$p\ r + (1-p)\ [pr - (1-p)\ (X+L)]>s\ (10)$$

or $p>1$ –[**Error!**]$^{1/2}$ (11).

Clearly, 0< **Error!**< 1, since r>s>0, and X, L• 0. Noting that p• 1 must hold in any event, we can derive the threshold $p^{;T}$ beyond which first movers would trust if $p^{;A}$ <p<$p^{;R}$ (for p>$p^{;R}$ they will trust anyway)

$p^{;T}$:=1 - [**Error!**]$^{1/2}$ (12)

Thus, for the parameter constellation $p^{;R} > p > p^{;A}$, first-movers will show trust if $p > p^{;T}$.

We are now in a position to describe fully the possible equilibria of the basic game of trust with courts, by reference to $p^{;R}$, $p^{;T}$, $p^{;A}$. With respect to the relative positions of these values and the population composition parameter p it may be helpful to consider Figure 5.

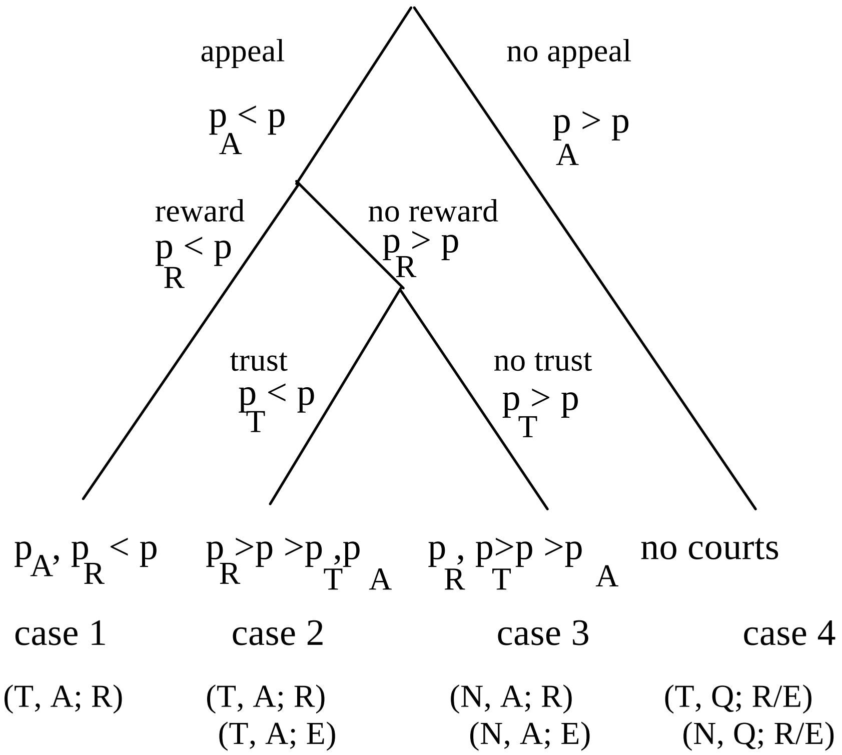


Figure 5

Before discussing the cases in some detail some basic observations may be helpful: First, if p > p ;R , p ;A all will trust, T, and reward, R, independently of their type regardless of any other relations between the parameters (case 1). Second, note that if p > p ;R , p ;A does not apply then either p • p ;A or p • p ;R or both are true (for, (p > p ;R p > p ;A ⁄(p • p ;A v p • p ;R)). Focusing on generic cases we consider again only strict inequalities. If p < p ;A then regardless of the location of other parameters -- in particular of p ;R -- courts will not be brought into play and nobody will show trust in the first place (case 4). Since p ;A > p is sufficient for case 4 to emerge we need to consider p ;R > p only for p > p ;A . If p ;R > p > p ;A either p ;T < p (case 2) with all first-movers trusting or p ;T > p (case 3) with no first-mover trusting emerges.

Case 1: p > p ;R , p ;A .

In this case the equilibrium strategy profile is: (T, A; R).

Reasoning: $p^{;T}$ is irrelevant since under rational play no exploitation takes place;

since $p > p^{;A}$, the threat of appeal by first-movers is credible;

since $p > p^{;R}$, the credible threat of appeal induces untrustworthy second-movers to choose R.

Case 2: $p^{;R} > p > p^{;A}, p^{;T}$.

In this case the equilibrium strategy profile is:

(T, A; R), if second-mover is trustworthy

(T, A; E), if second-mover is untrustworthy

Reasoning: since $p > p^{;T}$, first-movers will trust;

since $p > p^{;A}$, the threat of appeal by first-movers is credible;

since $p < p^{;R}$, this threat does not induce untrustworthy second-movers to choose R;

Case 3: $p^{;T}, p^{;R} > p > p^{;A}$.

In this case the equilibrium strategy profile is:

(N, A; R), if second-mover is trustworthy

(N, A; E), if second-mover is untrustworthy

Reasoning: since $p < p^{;T}$, first-movers will not trust;

since $p > p^{;A}$, the threat of appeal by first-movers is credible;

since $p < p^{;R}$, this threat does not induce untrustworthy second-movers to choose R;

Case 4: $p^{;A} > p$.

In this case the equilibrium strategy profile is:

(T, Q; R) if p>s/r and second-mover is trustworthy

(T, Q; E) if p>s/r and second-mover is untrustworthy

(N, Q; R) if p<s/r and second-mover is trustworthy

(N, Q; E) if p<s/r and second-mover is untrustworthy

Reasoning: since $p^{;A}$ > p and since courts are re-active this is basically the case with private type information and no courts. The discussion in section II of the case in which type detection is impossible directly applies.

When setting up a system of adjudication it is not beyond the influence of (constitutional) policy makers which of the four -- generic -- cases will prevail after the introduction of the courts' system. Though it is unlikely that X could serve as a policy variable it is quite plausible that L could be fixed as seems fit. Policy makers who seek to further 'the public interest' should choose L such that $p>p^{;R}$, $p^{;A}$. In that case (1) all players are led to behave in a trusting fashion and all will act so as to fulfil promises made. Moreover, this outcome is secured without the courts ever actually being brought into play. This is the force of the title of our paper: the 'shadow' of the courts suffices to generate universal compliance.

However, when fixing L policy makers face a trade-off: With increasing L the threshold $p^{;A}$ = **Error!**beyond which exploited first-movers would appeal increases while the threshold $p^{;R}$ = **Error!**beyond which untrustworthy second-movers would choose not to exploit first-movers' trust decreases. To maximise the range over which the conditions of case 1 are fulfilled, L must be chosen such that the maximum of the two thresholds $p^{;A}$ and $p^{;R}$ is minimised. Accordingly, set

$$p' := \min{}^{;L}(\max\{p^{;R}, p^{;A}\}) \quad (13)\ .$$

The interval (p', 1] is the maximum realm over which courts can conceivably influence p. Under optimal 'court policy' this realm is maximised. In the optimum we must have either $p^{;R} < 0 < p^{;A} = p'$ or $0\bullet\ p^{;R} = p^{;A} = p'$. This rules out cases 2 and 3 since both presuppose p ($p^{;A}$, $p^{;R}$). Moreover, since $p^{;A}$ is monotonically increasing in L, L=0 is the solution to the minimisation problem if $p^{;R} <0< p^{;A} = p'$ applies in a non-empty neighbourhood of L=0 (and thus over the whole range). Intuitively this makes sense, since with $p^{;R} <0$ no player in the role of the second mover will intentionally choose to exploit the first mover as long as the threat of appeal is credible. The latter is the case iff $p^{;A} <p$. Good court policy therefore suggests that the range of p for which players appeal be extended to its maximum; i. e. to set the policy variable L=0. If $0\bullet\ p^{;R} = p^{;A} = p'$ the value of L for an optimal court policy can be derived by solving $p^{;R} = p^{;A}$ or $3rX-(r+X) + (3r + 2X - 1)L + 2L^{2;}$ for $L\bullet\ 0$.

Under the court regime, if $p>p^{;A}$, $p^{;R}$, then everyone complies, and no cost is imposed on any player (except by mistake). Under the 'C, µ'-technology, not everyone complies -- except for the limiting values of C=0 and µ=1 --, and the cost C must be borne in every transaction. Provided that L can be fixed such that case 1 emerges, the courts' system will tend to secure a behaviourally better outcome at a lower cost than relying on the 'C, µ'-technology. This makes the introduction of systems of adjudication potentially attractive. But whether or not it would indeed be good policy to introduce a system of adjudication still hinges on the impact of the courts' system on the population composition. In evaluating the introduction of a system of adjudication it is not sufficient to point out the behaviourally superior results mentioned before. Somewhat deeper questions like the following must be raised as well: Can the court regime like the 'C, µ'-technology secure an equilibrium value of p that is sufficient to sustain the courts' benign operation? Will the court system have the (unintended) side-effect of 'crowding out' morally grounded dispositions and if so which are the relevant parameter constellations?

IV. Courts and the Population Composition

The court system operates in an environment in which type information is private. There are two basic ranges of p in the private type information case without courts: p<**Error!**with weak convergence and p>**Error!**with strong convergence. Accordingly we distinguish two classes of basic constellations in the shadow of the courts: that in which p $^{;A}$ • p $^{;T}$ • **Error!**and that in which p $^{;A}$ < p $^{;T}$ < **Error!**. Note that these two constellations are collectively exhaustive since

p $^{;T}$ < **Error!**⁄ p**Error!**< p**Error!**or p**Error!**• **Error!**⁄ p**Error!**• p**Error!**

(see the appendix for the simple derivation of these equivalencies). For population compositions p<**Error!**conditions with and without courts are identical if p $^{;A}$ > **Error!**. Therefore under the first constellation, p**Error!**• p $^{;T}$ • **Error!**, it suffices to consider the range p>**Error!**. Over this range the evolutionary dynamics of p as emerging under the influence of the courts must be compared with strong convergence of p towards **Error!**. Under the second parameter constellation, p $^{;A}$ < p $^{;T}$ < **Error!**, the focus must be on the range p< **Error!**. Over this range evolutionary dynamics of p in the shadow of the courts must be compared with weak convergence of p towards 0.

Assume initially that p>p $^{;A}$ • p $^{;T}$ • **Error!**. If L cannot be fixed such that p > p $^{;A}$ then case 4 which is equivalent to the basic game without courts emerges. The processes with and without courts are identical. If L can be chosen such that case 1 -- i.e. the parameter constellation p>p $^{;A}$, p $^{;R}$ -- prevails both types behave the same. Both act in a trustworthy fashion and the courts are not invoked. Under rational play nothing can differentiate between types. If there is any convergence towards **Error!**it must be mistake driven or weak. Thus, for p> p $^{;A}$, p $^{;R}$ • **Error!**the process in which p without courts declined 'swiftly' towards **Error!**must be slowed down if not stopped altogether by the presence of the courts.

Under optimal court policies -- as characterised in the last paragraphs of section III -- only cases 1 and 4 would have to be considered. Yet policy makers may fail to fix L optimally. Then (p $^{;A}$, p $^{;R}$)• cannot be excluded and either case 2 or case 3 could conceivably emerge. But case 3, p$^{;A}$ <p<p$^{;T}$,p$^{;R}$, is ruled out under the parameter constellation p$^{;A}$ • p$^{;T}$ • **Error!**. Only case 2, with **Error!**• p**Error!**• p**Error!**<p<p**Error!**, is possible. This case involves strong convergence to p$^{;A}$; and this for two reasons: first, since p$^{;R}$ > p, the expected pay-off of exploiting exceeds the pay-off to rewarding, so untrustworthy types in second-mover roles do better than trustworthy types; second, since p > p$^{;A}$, the courts are activated under rational play, and untrustworthy adjudicators do better (by an amount X• 0) than trustworthy adjudicators. Thus, even though in case 2 the courts may modify the pay-off structure as compared to the private type information case without courts, for p$^{;A}$ • **Error!**they cannot prevent strong convergence of p towards p**Error!**.

We can describe the relevant possibilities for p$^{;A}$ • p$^{;T}$ • **Error!**in terms of Figure 6. As in Figure 2, unbroken lines represent strong convergence; broken lines represent weak convergence. Within the category of weak convergence, we can make a further distinction between those regions in which the weak convergence depends solely on the evolutionary advantage of untrustworthy adjudicators (that is, on X), and those regions in which the untrustworthy benefit from mistakes for other reasons. The latter we denote in Figure 6 by double broken lines.

Since we are interested merely in generic cases we may assume that all parameters adopt different values. Taking into account that in all constellations presently under consideration we must have p$^{;A}$ • p$^{;T}$ • **Error!**there are, depending on the location of p$^{;R}$, only four strict orderings possible: **Error!**<p$^{;T}$ <p$^{;A}$ <p$^{;R}$, **Error!**< p**Error!**<p**Error!**<p**Error!**, **Error!**<p**Error!**< p$^{;T}$ <p$^{;A}$, p$^{;R}$ <**Error!**<p**Error!**<p**Error!**Since for p<p**Error!**the situation is equivalent to the situation without courts anyway we need to consider merely one of the three case in which p$^{;R}$ < p$^{;A}$ (see Figure 6a) . Figure 6b) shows the only remaining relevantly different ordering. In Figure 6 c) we show the (bench-mark) case in which type information is private and no system of adjudication exists.

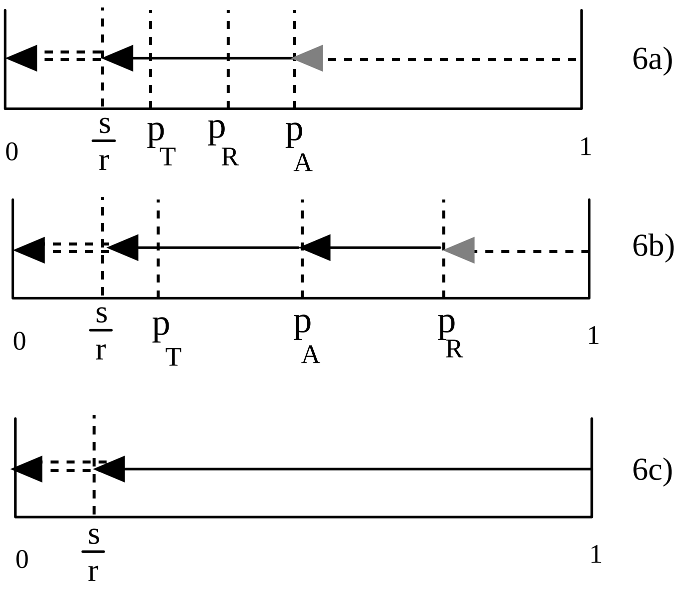


Figure 6 a-c)

Figures 6 a-b) show the intervals in which the courts can transform the strategy driven process of strong convergence of p towards **Error!**into a mistake driven process of weak convergence towards **Error!**. This happens iff the basic parameter constellation p > p $^{;R}$, p $^{;A}$ of case 1 prevails. Note also that the courts never operate to the strategic disadvantage of the trustworthy and for p > p $^{;A}$ • p $^{;T}$ • **Error!**slow down the decline of p. Thus, if p**Error!**• p**Error!**• **Error!**then introducing a court system is a dominant strategy for policy makers who seek to support T and R choices and intend to reduce the advantage of the untrustworthy.

Making policy recommendations would be easy if the constitutional strategy of introducing a court system would be dominant for p $^{;A}$ < p $^{;T}$ < **Error!**as well. Then under all conceivable parameter constellations introducing such a system would not favour untrustworthy individuals. However, for p $^{;A}$ < p $^{;T}$ < **Error!** the courts' system can conceivably accelerate the decline of p in some cases. To see which cases these are recall first (14), or p $^{;T}$ < **Error!**⁄ p**Error!**< p**Error!**. Thus we know that the ordering p $^{;A}$ < p $^{;T}$ < **Error!**must hold good. This reduces the number of possible cases to the four possible locations of p $^{;R}$ as shown in Figure 7 a-d). Again 7 e) shows the bench-mark case of private type information and no courts.

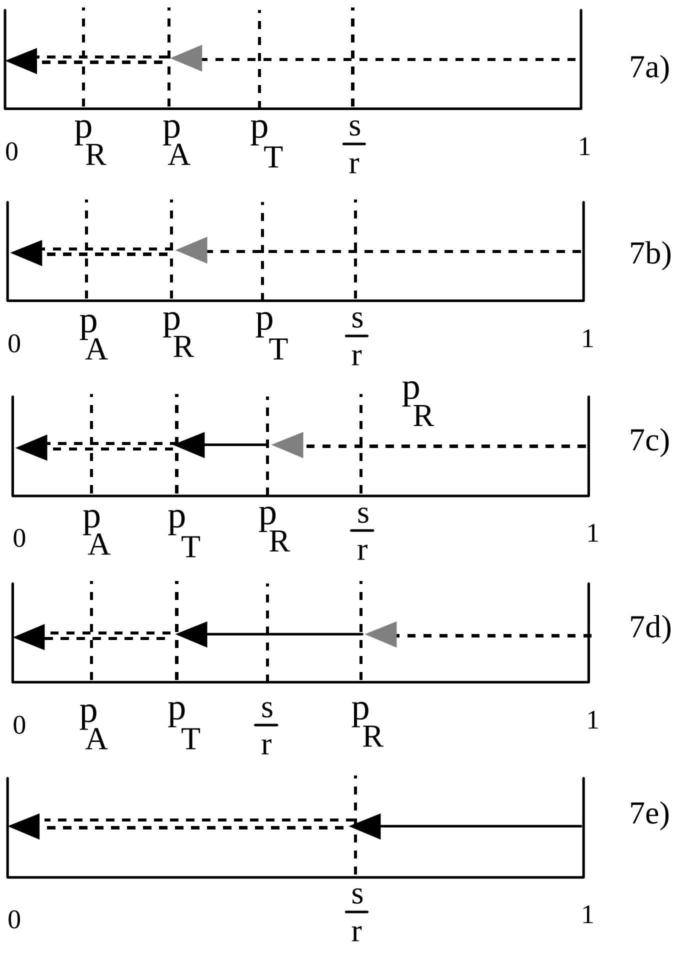


Figure 7 a-e)

If policy makers do not succeed in fixing L optimally, then $(p^{;A},\ p^{;R})$• , $p^{;A} < p^{;T} <$ **Error!**and p**Error!**>p**Error!**can emerge (see Figures 7c, d)). This is particularly relevant for p < **Error!**. For, in that case without courts no first-mover would trust. Yet the courts could conceivably induce first-movers to trust even though they eventually would be exploited by the non-trustworthy. If that happens the courts not only fail to slow down the decline of p but rather accelerate it and thus contribute to the crowding out of trustworthy individuals (see on crowding out Frey 1997).

There is obviously a generic interval $p^{;T} < p < p^{;R}$ for which the introduction of courts can harm the trustworthy. For example set X=L=0. Observe that

$p^{;T} <$ **Error!**⁄(r-s)(X+L+r) < r^2⁄X+L< **Error!**

This is certainly fulfilled for X=L=0 since 1>r>s>0 by assumption. Moreover, for X=L=0 the condition $p^{;T} < p^{;R}$ becomes

1 - [**Error!**]$^{1/2}$ < **Error!**.

Making s sufficiently small and choosing r sufficiently close to **Error!**, clearly, $p^{;A} < p^{;T} < p < p^{;R}$ can be fulfilled. Thus, in the case of sub-optimal court politics for $p^{;A} < p^{;T}$ < **Error!**there can be a generic interval in which the shadow of the courts actually works to the disadvantage of the trustworthy. Where in a situation with private type information no player would trust in the first-mover role now players by the presence of the courts are induced (or should one say 'seduced'?) to trust even though $p < p^{;R}$.

If optimal values of L cannot be secured then introducing the courts, though favourable under most parameter constellations and values of p is not a dominant strategy for policy makers who seek to slow down the decline of p. On the other hand, if L is chosen optimally then the courts never accelerate and often slow down the crowding out of the trustworthy. Though in Figures 6 and 7 all arrows point left, slowing down the decline of p may nevertheless be of fundamental value. Of course, how valuable such policies are depends on how much they reduce the potential advantage of the non-trustworthy.

So let us discuss weak convergence of p towards 0 and the adaptive process as it unfolds in the shadow of the courts. We shall take it that the probability of 'making a mistake' is type-independent. Often the consequences of such type independent mistakes do not affect types differentially. If, for example, a first-mover fails to trust when it would be rational to trust, then the second-mover receives s whatever her type. Similarly, if a second-mover fails to exploit when it would be rational to exploit, then both types receive the r pay-off. Such mistakes can not themselves induce a weak convergence process. They are neutral with respect to the evolutionary dynamics and thus may be left out of account when analysing the adaptive process.

Since the incidence of occasional lapses is type-independent and since the parameter m is purely subjective systematic differences in how different player types are affected by mistakes must depend on their strategic responses to mistakes. But as far as the latter are concerned only the untrustworthy can do differentially better -- either in their role as second movers in the basic game or as judges. Thus for non-neutral mistakes we must state: there is no mistake in any of the cases listed that positively favours trustworthy types. This explains in general why contrary to the case of the 'C, µ'-technology there are no arrows pointing right and no (limit) evolutionarily stable equilibria with p>0. Due to arguments like this one might infer that under fully rational behaviour the untrustworthy must always win in the 'long run' since when the 'golden opportunity' comes they will inevitably fare better. However, in our model this could conceivably be otherwise if external constraints on adjudicators' behaviour made it impossible for the adjudicators to get positive rents furthering their own evolutionary success.

Adjudicators operating under such constraints may be called 'judges' in the more narrow sense of that term which transports the notion of a person who does not have a stake in the case to be decided because her income does not depend on how she finds. As far as untrustworthy adjudicators might be in a position to set an L greater than the actual cost of litigation (L), but may not be able to appropriate that excess themselves they are in the role of judges too. In both cases the X as relevant for the evolutionary success of players in the role of adjudicators is driven to zero. Whenever this polar case (X=0) is feasible, then the courts could not only serve to ensure complete fulfilment of contracts irrespective of second-mover type, they could also sustain a positive population share $p>p^{;A}$, $p^{;R}$ of trustworthy types. The single dashed lines starting at the right borders of Figures 6 and 7 could be removed if X=0. In this single case there is no tendency, of either strong or weak form, for p to converge towards p=0. But there is no tendency to increase p, as well. Thus introducing a courts' system may generally be a good way to prevent the erosion of 'moral capital' but it will not actually build it up. For that to happen we must invest in more costly technologies like the 'C, µ'-technologies discussed in section II above.

V. Summary and Conclusions

Within the terms of our model of court process, adjudicators are not selected according to type but are drawn randomly at each turn of the evolving (market) process from the same population as ordinary players or traders who face the basic trust predicament. The central results indicate that adjudicative institutions *can* under plausible values of the parameters, and an appropriate initial proportion p of trustworthy persons, serve to secure three normatively desirable outcomes: first, that untrustworthy players are induced rationally to fulfil promises made; second, that in view of this fact all opportunities to engage in mutually advantageous trade can rationally be seized by an initial trustful move; and third, that at the same time the evolutionary forces leading to a disappearance of trustworthy types under optimal (constitutional) politics can at least be slowed, and possibly be halted altogether.

These results depend on having an initial population share p of trustworthy types that is not too low, but they do *not* depend on judges being any better on average than 'the rest of us'. The results also depend on some other premises which seem to us very reasonable. For instance, involving the courts as a device for solving the trust predicament in an environment where adjudicators would opportunistically exploit any powers they possess for their own purposes seems entirely to beg the question as to why the courts would reliably act to enforce contracts. Alternatively, if one is to assume that all adjudicators are 'trustworthy', then one ought on the grounds of symmetry assume that all players in the substantive game are similarly motivated, in which case one does not need the courts to achieve trustworthy behaviour in the first place. In other words, at either of the polar extremes along a notional motivational spectrum, introducing courts cannot be justified: they cannot do any normatively relevant work. However, what our results show is that at intermediate points along that notional motivational spectrum, courts *may* be able to add something important to contract enforcement, without any violation of the principle of motivational symmetry and without any assumption to the effect that type signalling or detection mechanisms are used to single out more trustworthy judges. This fact carries, we think, also an important methodological message -- namely, that in motivational matters at least, focusing on the polar extremes can be misleading. Intermediate cases can yield results that are not a convex combination of the results arising under the extreme assumptions (see Samuelson 1955, for a defence of the use of polar cases in another context).

Though our results may be welcome in particular within a Hayekian 'spontaneous market order' framework, they should not be misinterpreted. We do not explain how the adjudicative institutions of the market themselves might evolve 'spontaneously' -- only how they may conceivably work. Moreover, as long as we rely exclusively on adjudicators who can, if untrustworthy, draw a rent from their adjudicative activities, the introduction of the enforcement institution will not prevent a decline of the population share of trustworthy individuals. If no other forces work to their advantage the trustworthy will eventually be driven out of the population with concomitant effects on the scope and extent of the market. But interactions across markets are in general embedded (see Granovetter 1985) in a broader context of interactions in which due to the influence of a 'C, µ'-technology (or otherwise) the trustworthy do have some evolutionary advantage. Due to this embeddedness the effect of X>0, in particular if X is low, can presumably be compensated. But even if complete elimination of the advantages of untrustworthiness is impossible, alleviating evolutionary pressure on the trustworthy or eliminating the best niches for the untrustworthy is of great value and even where we cannot explain the emergence of virtue it is still worthwhile to organise our institutional life to 'economise on its presence'.

References

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 (1995). "Evolutionarily Stable Co-operative Commitments," *Humboldt University Discussion Paper- Economics Series* 53

Güth, W. and M. Yaari 1992. *An Evolutionary Approach to Explaining Reciprocal Behaviour in a Simple Strategic Game.*

Hume, D. [1739]1978. *A Treatise of Human Nature*. Oxford: Clarendon.

Samuelson, P. A. 1955. Diagrammatic Exposition of a Theory of Public Expenditure. *Review of Economics & Statistics* 37: 350-356.

Selten, R. (1988). "Evolutionary Stability in Extensive Two-person Games -- Corrections and Further Development," *Mathematical Social Sciences* 16(3): 223-266.

Derivation of (11)

$$p\ r + (1-p)\ [pr - (1-p)\ (X+L)>s \quad (10)$$

$$p\ r + (1-p)\ [pr - (1-p)\ (X+L)>s$$

$$pr + pr - (1-p)(X+L) - p^2r + p\ (1-p)(X+L)>s$$

$$2\ pr - (X+L) +p\ (X+L) -p^2r + p(X+L)-p^2(X+L)>s$$

$$2p(r+X+L)-p^2(r+X+L)>s+X+L$$

$$2p - p^2 > \textbf{Error!}$$

$$p^2-2p<- \textbf{Error!}$$

$$p^2-2p+1<1- \textbf{Error!}$$

$$(p-1)^{2<}1- \textbf{Error!}$$

$$(1-p)^{2<}\ 1- \textbf{Error!}$$

$$-p<-1 \pm[1- \ \textbf{Error!}]^{1/2}$$

$$p>1 \pm[\textbf{Error!}]^{1/2}$$

since p• 1 anyway, the condition is $p>1-[\textbf{Error!}]^{1/2}$ (11)

Derivation of $p^{;A} < p^{;T}$ ⁄$p^{;T}$ < **Error!**

$p^{;T}$ < **Error!**

⁄1 - [**Error!**]$^{1/2}$ < **Error!**

⁄1- **Error!**<[**Error!**]$^{1/2}$

⁄ (**Error!**)^{2}< **Error!**

⁄(r-s)(X+L+r) < r^{2} (*)

$p^{;A} < p^{;T}$

⁄ **Error!**< 1 - [**Error!**]$^{1/2}$

⁄ (X+L) < X+L+r - [(r-s)(X+L+r)]$^{1/2}$

⁄ [(r-s)(X+L+r)]$^{1/2}$ < r (both sides >0)

⁄ (r-s)(X+L+r) < r^{2} (*)

Note also (r-s)(X+L+r) < r^{2}

⁄ r(X + L)+ r^{2} - s(X+L+r) < r^{2}

⁄ r(X + L)- s(X+L)-sr < 0

⁄ (r-s)(X+L)<sr

⁄ X+L < **Error!**

Finally by negation ($p^{;A} < p^{;T}$ ⁄$p^{;T}$ < **Error!**)⁄(p**Error!**• p**Error!**⁄p**Error!** • **Error!**)