UNIVERSIDAD TORCUATO DI TELLA Escuela de DERECHO

Revista Argentina de Teoría Jurídica
Vol. 3, N° 1 (Noviembre de 2001)

TRUST, ADJUDICATION AND THE *QUIS CUSTODIET* PROBLEM

Geoffrey Brennan

Social and Political Theory Program,

Research School of Social Sciences, ANU

[September1998/ February 2000]

1. Quis Custodiet Ipsos Custodes?

There is a sense in which this familiar classical question - who shall guard the guardians? - is the central normative question in the public choice approach to politics and to institutional design more generally. It is so because public choice theory standardly makes what many would regard as extreme assumptions about the motives of those who hold political power - namely, that holders of political power will invariably tend to exploit that power to achieve their own ends at the expense of citizens at large. The assumption is nicely described by David Hume in a sentence often quoted in public choice circles:

"in assigning the powers of government and in devising the several checks and balances of the constitution, every man ought to be supposed a knave and to have no other purpose in all his action but self-interest". [Hume ' Of the Independency of Parliament' Essays Moral, Political and Literacy p117-118]

Actually, it is rather doubtful whether this sentence is representative of Hume's views

concerning either human nature in general, or concerning the assumptions of human nature relevant to political analysis in particular. But public choice scholars do not purport to be exegetes: they simply settle on the Hume statement because it represents their own position with a certain eighteenth century elegance. As Dennis Mueller (1989) puts it: " ....the basic behavioural postulate in public choice, as in economics, is that man is an egoistic, rational utility-maximizer." [p2]

Under this 'behavioral postulate' [more accurately a 'motivational postulate'], it is clear that the *quis custodiet* question is a serious one. It will be no problem to explain why a community will need to have rules and institutions for the enforcement of those rules: the rational egoism of individual agents will, under a variety of prisoner's dilemma predicaments, lead each to act in ways other than that which all would prefer. In principle, therefore, there will be an argument for rules that will ensure appropriate behaviour in such cases. But the very existence of the problem alerts us to the difficulty of finding a solution. If those rules are enforced by agents who are no different from those whose conduct the enforcement is to regulate, then who indeed will guard the guardians? Why will not the enforcers use their (necessarily) discretionary powers to simply advance their own private interests at the expense of everyone else? When every man is a knave, how is escape from a knavish society possible?

In this sense, the question, *quis custodiet ipsos custodes,* is seen by public choice scholars as a kind of presumptive knock-down challenge to any arrangement that involves the exercise of delegated power. And quite explicitly so. Public choice theory arose as a counter to the so-called "benevolent despot" model of government that inhabited policy economics - and more broadly the economist's theory of the state - as developed through the decades of the forties, fifties and sixties. The practice of applying normative criteria directly to policy parameters -as, for example, in the design of "optimal" taxes; or in deriving the ideal aggregate taxing and spending decisions for the fine-tuning of the macro-economy; or in combating instances of "market failure" in relation to public goods provision; or in achieving distributive justice (however exactly understood) - seemed to presuppose that the policy-adviser/economist was proffering her advice to a "benevolent despot". Both the despotism and the benevolence aspects of this conception were taken by public choice scholars to be objectionable - the despotism in that the policy-implementer was taken to be totally impervious to any political constraints, and the benevolence in that

the policy-implementer was assumed to be motivated solely by the pursuit of the "public interest" as articulated in the normative criteria so thoughtfully provided by the economist/advisor.

It was from the *quis custodiet* perspective specifically that the benevolence assumption was deemed to be so obnoxious. Not only did the benevolence assumption, at one stroke, set aside all *quis custodiet* concerns - what after all is the problem, if the guardians are totally benevolent? - but the assumption also implied that any intervention in the form of electoral competition or other constitutional constraint would only serve to prevent otherwise unconstrained policy-makers from doing the good they would otherwise do. In this sense, the benevolent despot approach seemed utterly undemocratic: it simply left no normative room in which political constraints might work.

Moreover, for economists, there was an utterly compelling argument against the benevolence assumption - namely, the argument from motivational symmetry. If 'market failure' was to be diagnosed on the basis of market agents acting in an entirely self-interested fashion, it would seem hopelessly biased ideologically to proclaim "political success" on the basis of agents acting in an exclusively benevolent fashion. What was required, observed the public choice critics, was an analysis of democratic political processes on all fours with the model of markets that allowed the diagnosis of market failure - with the "on all fours" requirement involving the same normative benchmarks and the same motivational assumptions as applied in standard micro-economic analysis of markets. Whatever else, public choice was an intransigent opponent of any assumption of motivational asymmetry: there was to be an insistence on the old folk-principle that politicians - and for that matter, bureaucrats and judges - are to be recognised as no better and no worse than the rest of us.

Of course, the principle of motivational symmetry could be met - just like the requirement of rationality - without any implication of egoism. Agents in both political and market roles could be modelled as somewhat benevolent; somewhat publicly-interested. Moreover, such a motivational structure might well seem manifestly more realistic than the pure egoist extreme. But what would still be ruled out is any normative argument for the exercise of government power that relies on the assumption that the rulers are in any way morally superior to, or more benevolent than, the ruled. Whether all normative arguments for delegated power depend on some such assumption is an open question. Does the quis

custodiet challenge represent the kind of knock-down challenge to delegated power under these more moderate motivational assumptions that it does in the extreme egoism case? That question is the one that this paper is expressly concerned to engage.

Before we turn to a more detailed analysis of this question, it is worth noting that the principle of motivational symmetry, though on its face reasonable enough, is not totally unobjectionable. Note, in particular, that it rules out the possibility of any effective selection of public officers on the basis of their 'fitness' for public office. Suppose, for example, that agents are not identical - that some are more motivated by the public interest, more benevolent or more naturally dutiful, than others. Then we might think of the 'civic virtue' that these more benevolent/dutiful agents possess as an asset which has particularly high social value in those employments where delegated power is to be exercised. We might imagine that selection procedures could be devised to assist both in the identification of such persons and the allocation of them once identified, to appropriate social roles. Perhaps in this spirit, for example, we might think of democratic electoral processes less as a way of providing incentives to candidates to offer policies in the interests of citizens (the standard public choice line) and more as a means of selecting to political office those with a strong sense of public duty. [Note that the principle of motivational symmetry would rule this possibility out]. In the extreme form, illustrated in the Hume quote, any form of selection according to differential civic virtue is ruled out because there is no motivational heterogeneity: "every man is supposed to be a knave". Note also that motivational heterogeneity in itself is not enough. We also need to have on hand adequately robust selection devices to distinguish the good guys from the bad. And this is no small challenge. After all, egoistic persons will want to occupy positions of delegated authority for precisely the same reason that we want especially publicly interested persons to serve - namely, that those positions provide scope for exploitation of others. Rational egoists will always have an incentive to masquerade as publicly interested persons in order to be appointed to positions where the discretionary power assigned can be exploited for their own ends.

The principle of motivational symmetry involves, then, either the assumption that all persons are motivationally identical or that, in the face of motivational heterogeneity, it is not possible to devise successful selection procedures. Certainly, the first of these possibilities is deeply implausible. We know, for example, from the emerging evidence yielded by a very wide range of laboratory experiments [see Sally (1995) for example] that individuals differ in their degree of public-spiritedness/benevolence - that, for example, somewhere between one third and one half of subjects behave “co-operatively” in social dilemmas (public goods provision experiments and the like), and that in simple division games different “dictators” distribute a given prize in differing proportions between themselves and the relevant claimant. The simple message that seems to emerge clearly and robustly from the entire range of these experiments is that agents are not totally egoistic, and that some are a good deal less egoistic than others.

Nevertheless, we might want to retain the assumption of motivational symmetry "on average". We might acknowledge moral heterogeneity among agents but think that devising reliable selection mechanisms that will distinguish good persons from bad is a hopeless dream. This, at least, is what I shall assume here. Throughout the argument that follows, I shall take it that there is motivational heterogeneity but that agents cannot *ex ante* distinguish the 'good' from the 'bad'. My aim is to show that, even in this somewhat unpromising soil, the principle of motivational symmetry does not rule out institutions of delegated power - at least under certain restrictions that I will try to spell out clearly. In other words, the *quis custodiet* challenge is not unanswerable. The challenge does not in itself without further argument prove that the *custodes* can do no good. In that sense, the cases in which everyone is totally benevolent or everyone is totally egoistic are equally misleading. In that same sense, the Humean analytic procedure for institutional analysis - the procedure that public choice orthodoxy routinely adopts [and which I have previously defended – for example, in Brennan & Buchanan (1985) ch.4] is unduly restrictive.

It may be useful at this point to set against the Hume quotation, a quotation from another authority figure - one that captures exactly the position on motivation that I believe is both right as a matter of fact and appropriate in the context of institutional analysis. The authority figure in question is Alexander Hamilton and the relevant observation comes from The Federalist Papers No. 76. Hamilton observes:

"....the assumption of universal venality is little less an error in political reasoning than the assumption of universal rectitude. There is a portion of honour among mankind that can be a foundation for our hope."

In this paper [as well as in other recent work – such as Brennan & Hamlin (2000)] I want to argue exactly along these Hamiltonian lines.

2. Trust and Contract:

The general argument to be advanced here I will develop in terms of the 'rational actor' analysis of trust. I want specifically to examine the implications - if any of that analysis for the role of the courts in enforcing contracts entered into and/or promises given. The analysis is not intended so much as a piece of law and economics as it is a piece of political theory. I shall therefore not be concerned with the precise content of the laws. Nor shall I be concerned to model the institutional structure of the courts in the most plausible way. Indeed, the picture of the courts I shall offer is so remote from any prevailing legal institutions that I prefer to talk of institutions of "adjudication", as in my title, rather than of courts as such. However, the institutions of adjudication are taken to be such that once adjudication is involved the adjudicator's decision is enforced: the *custodes* in the account I shall develop do have genuine powers. The particular assumptions I have made about the courts are designed to accommodate easily the principle of motivational symmetry – not necessarily to mirror reality.

My primary object in this section is to introduce the social predicament that the "economics of trust" exposes, and to discuss briefly certain possible "solutions" to that predicament. The whole treatment here will be discursive and illustrative. The conclusions do derive from a more technical treatment [Brennan, Guth & Klient (1992)] but the more elaborate analysis is not necessary or making the central points. My object here is to argue for those conclusions intuitively and in a way more explicitly focussed on the political theory implications.

The analysis of trust has recently become something of a minor industry in academic social sciences. The last few years have seen a number of books and countless articles on the topic and on related questions about the erosion and maintenance of "social capital". Economists have not been exempt from this fashion - and for good reason. Generalised trustworthiness is part of the oil of commercial society: trust makes the wheels run more smoothly (it lowers transactions costs) and it makes possible certain mutually beneficial exchanges that would otherwise not occur. Indeed, some (possibly small) measure of trust is arguably necessary in all economic transactions: someone must always move first in any economic transaction and thus "trust" the second-mover to fulfil on the bargain.

For the economist, the analysis of trust and trustworthiness must be nested within a broad rationality framework -- though as we shall see, in some ways the phenomenon of trust represents something of a challenge to that framework.

In the standard economic account of trust, all action on the trusting side is driven by rationality: it is on the trustworthiness side of trust of trust relations that the challenge to rationality appears and on this trustworthiness side that most of the intellectual interest focuses. Consider, specifically, the basic trust game, designated Fig.1. It is a two-person sequential game, depicted here in extensive form. Player I moves first and must choose between not-trusting (N) and trusting (T). If I chooses N, the game ends and both players received a pay-off of zero: pay-offs are denoted at each node by a number pair (*a*,*b*) where *a* is the payoff to player I and *b* the pay-off to the second-mover, designated II. If I chooses T, than II gets to choose - between options of exploiting (E) where I gets a pay-off of -1 and II gets a pay-off of 3 and rewarding (R) where both receive pay-offs of

2. All these pay-offs are common knowledge - that is, they are known by both players and known to be known.

In the trust game as depicted, there is an equilibrium for rational players -namely, N. For I knows that II, if rational, will choose E over R, thereby receiving a pay-off of 3 rather than 2. But under E, I receives -1 whereas he would receive zero under N: hence I will not trust. And this reveals the predicament. Like the prisoners' dilemma, this game has an equilibrium that is 'Pareto-dominated' by another technically feasible outcome, R. That is, the equilibrium at N involves lower pay-offs for both players than each would obtain at R, but R is rendered inaccessible by virtue of II's rationality.

[Perhaps it is worth pointing out here that by attributing the inaccessibility of R to II's rationality, I am committing myself to a particular view of rationality (the standard economist's view) that is philosophically contentious. The issues at stake here are worth noting. If it is true that rationality in the arena of action makes one's life go less well than it would if one were identifiably "irrationally" trustworthy, then one has rational reasons to become (irrationally) trustworthy. It is clear that our terminology is inadequate here. We need to distinguish A-rationality (rationality in the arena of action) from D-rationality (rationality in the arena of dispositions, where a disposition is to be understood as a mode of calculation that might weigh attributes other than the objective payoffs in deciding action). What we might then say is that it is D-rational not to choose the disposition of A-rationality. It is I think, unhelpful to argue (as Gauthier does) that if player II forms the intention to choose T in the first period, this intention formation makes R the rational choice of action in the second period. It seems to me that it is, at best, misleading to say that R is the "truly rational" option for II to choose *because the intention to R was rationally formed*. I want to insist that the rational action for II is E (siding, as I understand it, with Derek Parfitt on this question). In the same spirit, I would want to reject the suggestions of Elster (1981; 1983 for example) that the problem arises here because of imperfect rationality. This description suggests that if II were fully rational, II would choose R not E. In my view, it is a characteristic feature of rationality that rational choice is forward-looking and unrelievedly expedient: only if E were not the better choice from the position of the second-chooser when actually located at the choice node 2 could we say that R was the rational action.]

Fig.1: The Basic Trust Game

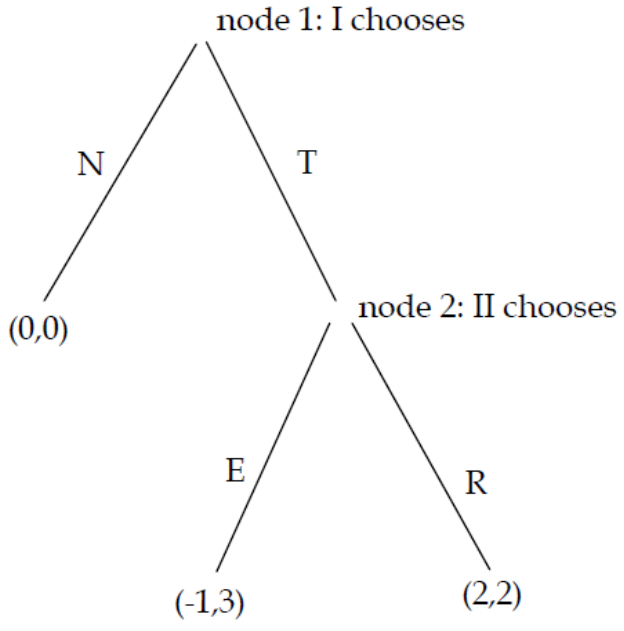
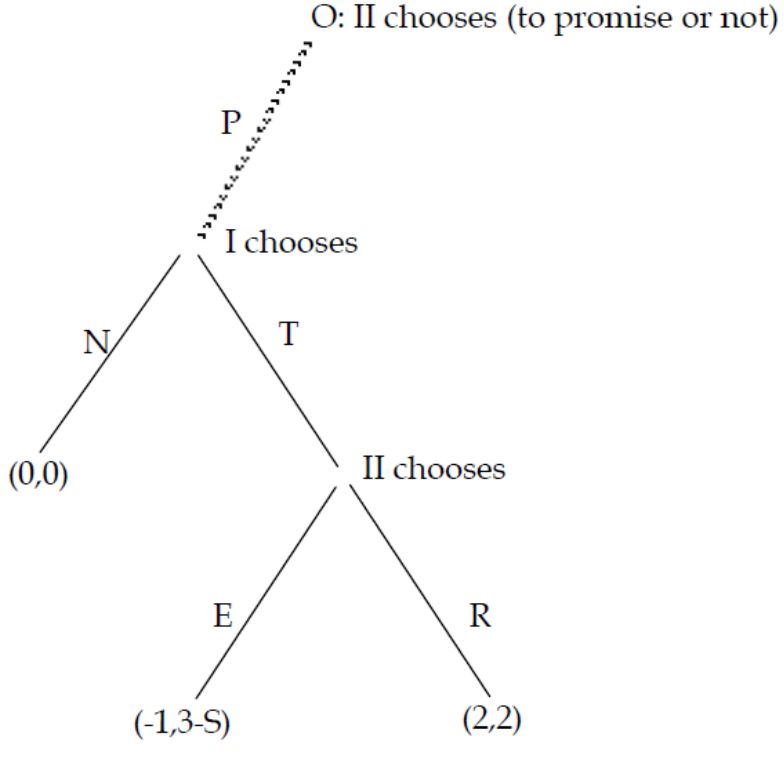


Fig.2: The Basic Trust Game with a Trustworthy Second-Mover



Of course, if Fig.1 were not a complete depiction of the interaction between I and II - if, for example, the interaction is ongoing and II's choosing E in the current play induces I to choose N in the future, then there may be reason for II's choosing R. But this point, familiar to economists as 'the discipline of continuous tradings', though entirely valid and useful in itself, is beside the point in the trust game. Even if the trust game is replayed, and players' reputations are publicly available, there will always be some interactions that have the structure of the game as depicted: there will always be some, potentially mutually beneficial, exchanges that will not be consummated by fully rational actors (because rational action for the second-mover will dictate choice of E).

The notion of D-rationality does, to be sure, suggest a solution to the trust predicament - a 'solution' that brings us closer perhaps to our intuitions about trust. It may, we have said, be rational at the level of dispositional choice for II to choose a disposition to keep promises. For such a person, bygones are not bygones: a promise given, a commitment made, will, we might say, be costly to break. We will characterize such a trustworthy person as being one whose payoffs are those shown in Fig. 2 where *S* is a 'conscience parameter' and takes a value in excess of 1. If we want to maintain the spirit of instrumental forwardlookingness, characteristic of rationality, we might think of *S* as the burden of shame and guilt that II will endure if II fails to fulfil a commitment given (we should emphasize however that *S* does not have to be constructed in that instrumental way). (In fact, Fig.2 is a truncated form of the actual game: to be complete, we would need to show an earlier (period 0) choice by II to promise or commit (strategy P)).

However, it is a crucial feature of D-rationality that it will only pay II to choose the trustworthy disposition if doing so increases the chance that I will trust him. If I trusts "blindly" - independent of whether he has reason to believe that II is trustworthy or not - then untrustworthy second-movers will do better than trustworthy ones: it will not then be D-rational to be trustworthy.

Consequently, a crucial factor in the (D-)rational trust story is the degree of "translucency" in relations among persons - the extent to which first-movers can identify second-movers dispositions. To take the limiting case, if first-movers can identify trustworthy second-movers perfectly, then it will pay all agents to adopt the trust-worthiness disposition.

Or at least, this will be the case if dispositions can be rationally chosen in this way. The idea of D-rationality simply presupposes that choice of dispositions is freely available - that we can reliably acculturate ourselves to keep commitments. But there may well be difficulties here. Suppose, as D-rationality would seem to require, that the reason we try to acculturate ourselves to the disposition of keeping our commitments is that we think we will be rewarded thereby by being trusted. Now, however, precisely that same consideration will give us reasons for *not* keeping our commitments when we come to choose whether to act in a trustworthy fashion. If we are genuinely to feel shame and guilt for breaking our promises, it must be because we think breaking promises is wrong - not because we think that thinking breaking promises is wrong will be good for us. An example from John Broome is telling in this connection. Suppose you are sick. And suppose your rate of recovery depends to some extent on your state of mind. Specifically, let it be the case that, if you believe that you will recover in n days, you will recover in (n+2) days where n 1. Clearly, it is best for you to believe that you will get better tomorrow, because then you'll get better in three days time. But the desire to get better as soon as possible can't provide a reason for the belief that you will get better tomorrow: the only proper reason for you believing that you'll get better tomorrow is that you happen to think that the relevant proposition (that you'll get better tomorrow) is true. It might well be that trustworthiness is like that. Perhaps the only way to be trustworthy is to have a genuine belief that breaking undertakings is morally wrong. And beliefs of this kind cannot simply be prudentially manufactured. Or so it might be argued. Whether the disposition of trustworthiness hangs on certain moral beliefs, and whether such beliefs are inaccessible to rational choice are issues that we need to note, but cannot here hope to answer. For my part, I am inclined to think that there probably *is* an irreducible moral component embedded in feasible dispositional "choice". However, even if this were not so, having the relevant moral beliefs would certainly seem likely to make your trustworthiness more robust.

If trustworthiness is D-rationally inaccessible in this way, we might nevertheless provide an account of the emergence of trustworthiness in evolutionary terms. Evolutionary accounts do not need to make any appeal to agent intentionality. We would certainly expect that trustworthy types will do better than merely (A-)rational ones, supposing that doing better in pay-off terms correlates well with survival characteristics. Even here, of course, it remains necessary that agents' dispositions be sufficiently 'translucent' in the system for trustworthiness to be evolutionarily advantageous: first-movers must be able to tell trustworthy and untrustworthy types apart with some degree of accuracy. Otherwise being trustworthy won't increase payoffs. But what if the required degree of translucency is absent? What if, just as our powers to select good guys as judges is inadequate, so our capacity to select good guys as trading partners is imperfect? What, if anything, can we do?

3. Enter the Courts, With Praise:

It is perhaps natural that anyone coming to the trust predicament - at least anyone not deeply acculturated into public choice logic - will see that predicament as the central rationale for the law of contract. Precisely because agents cannot make credible commitments in lots of cases where it would be profitable for them to do so, we need the law of contracts, backed by enforcement powers, to make those commitments credible. If, however, we apply public choice logic - and the principle of motivational symmetry in particular - then we immediately confront the *quis custodiet* problem. If the judges and police are taken to be fully trustworthy, then so must ordinary traders be . And if ordinary traders are fully trustworthy, we do not need the courts. Alternatively, if traders are routinely untrustworthy, then to intrude the courts into the trust predicament is simply to admit another player, with special powers, who is every bit as rationally expedient as all others. There does not seem to be any *a priori* reason why this third player would reliably enforce contracts between the other two. Surely she will simply expropriate all available rents within the limits (if any) of her assigned powers. In these two cases, at either extreme of a notional motivational spectrum, the courts appear to add nothing. They can do no useful work if everyone is trustworthy (because we don't need them) or if everyone is untrustworthy (because judges won't reliably enforce the law). But that is not the end of the story. In at least some intermediate cases - cases in which there are both trustworthy and untrustworthy types around - the courts, understood as institutions of adjudication with genuine enforcement powers, can indeed do positive work. Just what that positive work is I shall shortly specify.

To clarify the kind of assumptions about knowledge of other players that I have in mind here, consider an evolutionary context in which there is initially a proportion P of trustworthy types, where $O < P < 1$. Players are selected at random from the population to play the basic game and their roles (as first or second mover) are also determined randomly. It is assumed that P is known by all players, so each first-mover knows that the second-mover will be trustworthy with probability P. In order to ensure that the principle of motivational symmetry obtains, adjudicators are selected randomly from the same pool as players; accordingly, the probability that an adjudicator will prove trustworthy is that same probability P. Players play the trust game once, and then are reallocated for a further random match in subsequent games. Players are anonymous, so there is no possibility of

building a reputation.

I am here going to make a critical assumption about the role of adjudicators. This is that the adjudicator role is reactive in the sense that adjudicators can only be called into play by one of the players in the substantive game. So, while adjudicators have genuine power to have their judgements enforced and though this power will enable them to collect rents from the players, the adjudicators only get to exercise this power by invitation.

How will adjudicators behave? I take it that trustworthy adjudicators will decide for the exploited party, if there is one, and enforce a redistribution from the exploiter to the exploited to ensure that the exploited receives what she was promised. So, in our original trust game for example, if the second-mover chooses E and the adjudicator is activated and proves to be a trustworthy adjudicator, then the adjudicator will transfer 3 from the second-mover to the first-mover, to ensure that the first-mover receives the pay-off of 2 that was promised; this will leave the second-mover with a pay-off of zero. I take it, further, that untrustworthy adjudicators will decide cases randomly; but will systematically appropriate rents from the players, an amount X from player 1 and an amount Y from player 2. So if the outcome is E and the adjudication process is activated, in the case where the adjudicator is untrustworthy, the payoff structure will be:

11

Ø ø
(2 -1)-*X* = -*X* for I

Œœ

2 º 2 ß

and

1 Ø 3 ø

(3 + 2) -*Y* =-*Y* for II

Œœ

2 º 2 ß

Now, there is an important simplification at hand here. Since there is always a chance (1-P) of having an untrustworthy judge, there is an expected cost to the players in invoking the adjudication process. Note that an exploited first-mover is the only player who can ever gain anything from the adjudication process. Hence the only case in which the adjudication process will be activated is when E is the outcome and the invitation comes from exploited first-mover. Of course, if X and Y are large relative to the pay-offs (and there is nothing yet to indicate why they should be unbounded), even an exploited first-mover may not activate the adjudication process. But the simplification here is that, in order to determine whether the adjudication process would be invoked or not, we need only examine the calculus of an exploited first mover.

On this basis, we can treat the trust game in the presence of the courts as an expanded version of the basic game. Effectively, what happens is that if E were to emerge as the outcome of the game, the first-mover, I, gets a further choice: he can either be quiescent, Q, and accept the (-1,3) outcome; or he can activate the adjudication process, V. In the latter case, the outcome is (2,0) with probability P

(13)

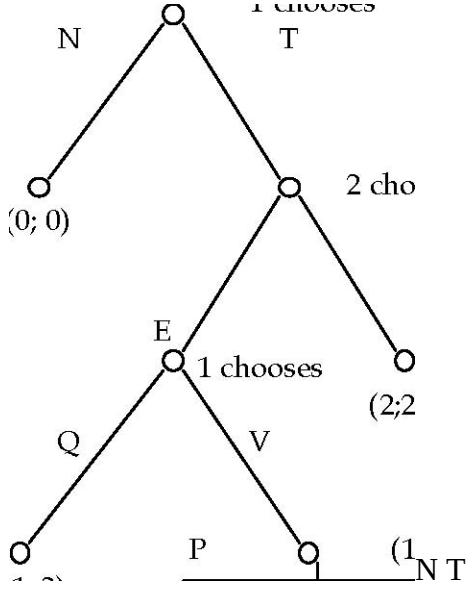
(i.e. when one gets a trustworthy judge) and $^{-X,-Y}$ with probability (1

Ł22 ł

P) (i.e. when the judge is untrustworthy).

The entire game tree is laid out in Fig .3 (although we show the option of invoking the adjudication process only for the case where the second-mover is untrustworthy and chooses E). The contribution of the adjudication process can then be assessed by a comparison of the game with the process and without it.

Fig. 3



1 chooses

2 chooses (0; 0)

ER 1 chooses (2;2) QV

P (1-p) (-1; 3) (2; 0) (1/2 – X, 3/2 – Y)

Consider the case without the courts first.

Player I will reason that player II will prove trustworthy in P cases and

untrustworthy in (1-P) cases. The expected return from choosing T is therefore:

(T) = 2.P + (-1) (1-P) = 3P - 1

So, I will trust if P > 1/3and not trust if P < 1/3

1

The game pay-off is therefore: for P > (3P-1,3-P), total 2+2P

31

for P < (0,0), total zero

3

Note, further, the effect on the “evolutionary equilibrium” value of P. If the

1

initial value of P is P1> , then I will trust. Accordingly non-trustworthy types

3

in second mover roles will do better than trustworthy types in second-mover

1

roles; so P will decline, until we reach a point where P< , where no-one will

3

trust. At this point, trustworthy and non-trustworthy types do equally badly

and P will decline no further. Over this range, the fact that there are trustworthy

types around does no good: there simply aren't enough such types to induce rational first-movers to trust, and the pay-off to all players in the game is zero.

The case in the presence of the courts is somewhat more complicated. The central issue here is whether and under what circumstances, I will choose V rather than Q. For if it is rational for I to choose Q, II will know that fact, and the solution to the extended game will be identical with that in the case without adjudication. Accordingly, we focus on the first-mover's choice whether to invoke the adjudication process (V) or not (Q). The first-mover will choose V only if:

$$(\tfrac{1}{})$$

$$2P + (1 - P) - X$$

$$Ł_{2} ł^{\geq -1} \quad \underline{2X - 32X - 3}$$

i. e. $P \geq$. *DenotePv* =

$$2X + 32X + 3$$

This parameter, Pv, represents the threshold value of P, such that if P > Pv, then I will appeal. For illustrative purposes we might note that Pv is rather less than 1 even for quite sizeable values of X relative to pay offs in the basic game: for X=8.5, P=.7; for X=10, P=.75; for X=3.5, P=.4. Obviously, as X becomes extremely large, then Pv becomes closer and closer to unity, as we would expect. If the rents that an untrustworthy adjudicator can extract are very large, the proportion of trustworthy persons required for I to rationally invoke the adjudication procedure will have to be correspondingly larger.

At this point, note what happens if this condition is satisfied. Player II will know that this is the case. And player II can act to avoid the outcome V by choosing R

$$\underline{1}$$

in the earlier period. Will it be rational for him to do so? Yes, if 0.p + (1-p) (1

$$2$$

Y) <2. And this condition will always be satisfied under plausible values of Y.

$$\underline{1}$$

(In fact, in this example, we require only that Y be greater than) In this sense,

the scope for an untrustworthy judge to secure significant rents is an advantage, because it induces untrustworthy players to fulfil their commitments so as to avoid having to face the courts. If the proportion of trustworthy adjudicators is high enough to induce the exploited first player to invoke the adjudication procedure, then untrustworthy second-movers will rationally act to make such invocation unnecessary by *acting* in a trustworthy manner.

Consider then the equilibrium in the game with the adjudication procedure where P > Pv. Here, everyone - trustworthy and untrustworthy alike - will fulfil commitments. Final-movers will know this (knowing P) and hence will trust. So the outcome is:

Moreover, for cases where the initial value of P exceeds Pv, trustworthy and untrustworthy types are behaviourally equivalent, so in particular untrustworthy types have no evolutionary advantage. There is therefore no reason to think that trustworthy types will be driven out in the evolutionary equilibrium. Of course, trustworthy types do much less well than

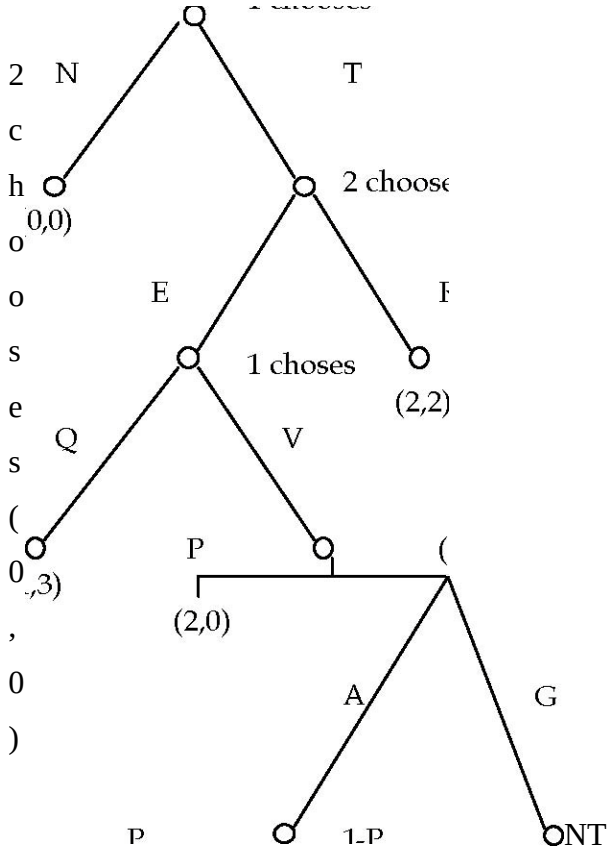
for P > P_v	:	(2,2)	total 4
for P_v > P> 1 3 :		(3p-1,3-p)	total 2+2p
for P < 1 3	:	(0,0)	total 0

untrustworthy types in their roles as *adjudicators*; but the peculiar genius of this adjudication arrangement is that the adjudication process is never actually invoked. Because the adjudication process imposes asymmetric net costs on untrustworthy players, they find it expedient to keep their commitments, precisely to avoid facing adjudication. Adjudicators never get to act, so the payoff differential to untrustworthy adjudicators is irrelevant.

Obviously, a critical question in this entire story is how stringent a hurdle Pv represents: if Pv is reasonably close to unity, then the adjudication process can only offer us any advantages in rather extreme conditions where P is high. This critical value, Pv, depends crucially on X, the rents that an untrustworthy judge can extract from first-movers. What value of X is plausible? In particular, how can X be bounded?

Our construction itself suggests a mechanism by which X can be bounded. Suppose we impose a further layer of adjudication on the process laid out so far: that is, we allow substantive players to appeal the decision of the first-order court. At this appeal level, however, there is no further constraint, so we might expect that untrustworthy adjudicators at this level will appropriate truly maximal rents. Suppose the maximum that can be extracted from the players is

M. We can depict the situation in game-tree form in Fig.4. This is simply Fig.3 with the extra option for I (who is most relevant here) to appeal the decision should he get an untrustworthy adjudicator in the first-round adjudication process. The options are either appeal [denoted A], or accept the verdict [denoted G]. Note that the untrustworthy first-round adjudicator will expect to lose in the event of appeal, whether the second-round adjudicator is trustworthy or not. If the second-round adjudicator is trustworthy, he will correct the first-round decision, restore I's promised amount of 2 and restore the mis-appropriated rents obtained by the untrustworthy adjudicator to their original owners. If the second-round adjudicator is untrustworthy, he will appropriate all available rents, not only from I and II *but also from the first-round adjudicator.* (The third-terms in the pay-offs in the final pay-offs from period 4 indicate the pay-offs, to the first-round adjudicator). The critical point here is that it will not be rational for that first-round adjudicator to set X and Y at levels that would induce I and II to appeal: for if there is an appeal, that adjudicator will lose.



ER

1 choses (2,2) Q V chooses

P

(1-p) 1 chooses (-1,3) (2,0)

AG

P 1-P (1/2 – X, 3/2 – Y) (2,0) (1/2 – M, 3/2 – M)

[M>x,y]

Accordingly, it must be the case that: 2p + (1-p) (1/2-M) < 1/2 –X for I not to appeal

1

and 0.p + (1-p) (3/2-M) < 1 - Y for II not to appeal

2

So the condition on X becomes:

1

M -X +

X

2

p < which is roughly < 1-p

M + 2 M

And the condition on Y becomes:

Y 5

< 1-p+

M 2M

Note that the condition is more stringent on X than on Y, so there are reasons to think that untrustworthy judges, subject to possible appeal, will extract larger rents from exploiting second-movers than exploited first-movers. More to the point perhaps, the rents taken by a rationally untrustworthy first-round adjudicator will be bounded, perhaps to one third of the total appropriable rents, so as to ensure that it is not rational for players to appeal.

4. Conclusion:

This piece of the argument completes the picture. The institutional arrangement on offer consists of randomly selected adjudicators drawn from the same population as players, with an appeal mechanism (involving a further randomly selected adjudicator) available to players. The adjudicator's judgement is enforceable: adjudicators have genuine power to that extent. But the adjudication process cannot be activated by anyone other than the players themselves. The process is "reactive" in that sense.

What this procedure delivers is an outcome in which all players fulfil contracts, for reasonable values of the proportion of trustworthy persons - values, certainly, considerably less than 100 percent. Moreover, the process serves to prevent the erosion of trustworthiness in circumstances where trustworthiness *would* otherwise erode. These accomplishments are major ones. It is true that securing those accomplishments depends on the proportion of trustworthy persons in the society not being too low - but it does not require men to be angels. Well short of universal rectitude, to re-echo the earlier quotation from Hamilton, there can be a portion of honour among mankind that can be the foundation of our hope.

For those without a taste for arithmetic, it might be useful to state what is at stake in the foregoing argument, shorn of all the analytic detail. Simply put, the story is this. The trust problem arises because parties who act rationally cannot make credible commitments – there is no available means whereby a person who chooses at every point that action that makes life go best for him can fulfil promises given. What the courts, under the description given here, provide is an effective mechanism for first movers to make credible threats – at least under certain circumstances. The reason the courts provide this facility is that first movers who are exploited – who trust someone who fails to fulfil undertakings made – can appeal to the courts; and the threat of that appeal will induce rational second movers to fulfil. This result will follow even if courts are somewhat corrupt, provided that they are not *totally* corrupt – that is, provided that not *every* man is a knave. It requires however that courts are *reactive* – that they are activated only by appeal from a party to the substantive contract. It does not require, on the other hand, that agents be able to distinguish trustworthy from untrustworthy types, either in the choice of trading partners or in the selection of persons for the adjudicatory role. In that sense, it does not involve any violation of the principle of motivational symmetry: judges are no different (and specifically no

better) than the ordinary players. In this sense also, the question as to who shall guard the guardians is seen to be a second order question – and certainly not necessarily a knockdown challenge to institutions of delegated power. In the model presented, the courts can do normatively relevant work even when the guardians are not guarded. The creation of their presence, in and of itself, can be sufficient.

There are two particular comments on details of the model - two possible criticisms - of which I am aware and to which I have no satisfactory answer, at least at this point. One relates to the nature of trustworthiness assumed. I have here treated trustworthiness as a disposition. That is, *S* is large enough to ensure trustworthy behaviour from a trustworthy type irrespective of temptation. This means that the proportion of trustworthy adjudicators is identical to the proportion of trustworthy players, despite the fact that agents in adjudicator roles face more substantial temptations to exploit than those same agents do in their roles as ordinary players. I concede that this assumption is implausible. I am inclined to believe that there is, as we might put it, a downward-sloping demand curve for morality as for other things. The way in which I have modelled dispositions is, in this sense, excessively rigid. It would be possible in principle to allow p to be a function of the stakes at issue, but doing this would make the analysis much more complicated and would not, I believe, alter the qualitative results.

The second criticism relates to the assumption of reactive courts. Who, one might ask, is to enforce this requirement? Does not this requirement return us to the *quis custodiet* problem yet again? Possibly. My response here is merely to observe that some institutional arrangements make greater demands on human virtue than others. Economists have a professionally grounded inclination to think that any demand on human virtue *at all* is an automatic signal of infeasibility: this is the message of the Hume quotation cited at the outset. But I am siding explicitly with Hamilton here. As I have observed elsewhere, one does not have to believe something is *infinitely* scarce to believe that it is worth economising. As the model here illustrates, if one takes a less stringent (and in my view more plausible) line on human virtue - a line closer, say, to Adam Smith than to Bernard Mandeville - then possibilities for institutional design offer themselves which in the more extreme motivational picture will be simply ruled "out of court" as we might put it.

Finally, while we're on the question of matters "out of court", it is an interesting feature of the picture I have sketched that the courts are never actually invoked. Courts do their work by creating credible threats, rather than by action. And it seems to me that this result is likely to be robust in any model in which rational action is assumed. That is, in models of court activity with essentially rational players, it is likely to be a feature that the courts will never be invoked except by accident – and this under a much wider range of institutional details than those sketched here. If the courts' primary role is to alter incentives to secure particular outcomes, then rationality ought to ensure that court procedures are not invoked. In this sense, the fact that courts are 'reactive' in the terms used here is an extremely important feature of the whole argument. A general defence of appropriate rules of 'standing' follows more or less directly from the foregoing argument.

BIBLIOGRAPHY

Brennan, Geoffrey & James Buchanan (1985) *The Reason of Rules*, Cambridge

University Press, Cambridge. Brennan, Geoffrey & Alan Hamlin (2000) *Democratic Devices and Desires* Cambridge University Press, Cambridge.

Brennan, Geoffrey, Werner Guth and Hartmut Kliemt (1998) 'Trust in the

Shadow of the Courts' [mimeo] Hume, David (1985) *Essays Moral, Political and Literary* (ed E. Miller) Liberty Press, Indianopolis.

Madison, James, Alexander Hamilton and John Jay (1788/1987) The Federalist Papers (ed. I. Kramnick), Penguin Books, Middlesex.

Mueller, Dennis (1989*) Public Choice II*, Cambridge University Press, Cambridge. Sally, David (1995) 'Conversation and Co-operation in Social Dilemmas' *Rationality and Society*, p58-92.