CONSTITUTIONAL POLITICAL ECONOMY, VOL. 3, NO. 2, 1992

THE EMERGENCE OF A PROTECTIVE AGENCY AND THE CONSTITUTIONAL DILEMMA

Ulrich Witt*

In the last decades' revival of contractarianism a constitutional contract is interpreted as a device to overcome the hypothetical state of anarchy. It is not entirely clear, however, how, in a pre-constitutional setting that lacks any institutional forms, an unanimous agreement on the rules and the agency enforcing the rules can be imagined to emerge. This paper conceptualizes the problem in game-theoretic terms. A solution is discussed together with an old dilemma that turns up in this context. The dilemma results from the fact that the protective agency has to be endowed with sufficiently powerful coercive means to prevent anyone breaking the social contract. However, this concentration of power itself may induce a violation by making the protective agency usurp its power. The logical basis of the dilemma is explored together with the conditions under which it may challenge the contractarian approach.

Introduction

The legitimacy of the state is a central theme in moral and social philosophy. It has been given a new lease of life by the work of Buchanan (1975), Nozick (1974), and Rawls (1971), the "new contractarians" (Scott 1976). In their work, which revives a tradition reaching back to Locke, Rousseau, and even Hobbes, the basic idea is to postulate a hypothetical unanimous consent which legitimizes the 'protective state', i.e. the authority enforcing and protecting individual rights in social interactions. In an imagined state of anarchy ('Hobbesian jungle', 'state of nature') all members of society are supposed to decide on a suggested constitutional contract. The contract would determine the rules of orderly interaction within society after unanimous consent has been given and would establish an agency to enforce those rules. It is argued that, given certain preferences of the members of society, such a contract, and the agency enforcing it, would be approved as they would enable society to escape permanent anarchy and the Pareto inferior outcomes associated with it.

*Professor of Economics; University of Freiburg, 78 Freiburg, Germany. I would like to thank Juli Irving-Lessmann, Dennis Mueller, Viktor Vanberg, Georg von Wangenheim, and in particular Hartmut Kliemt for helpful comments on an earlier draft.

In a large body of literature many facets of this basic idea have been elaborated in detail. Yet one problem, hidden behind the catchword of the "invisible hand explanation" of the emergence of a protective agency (Nozick 1974: chap.2), has found comparatively little attention: the question of how precisely, in a pre-constitutional setting that lacks any institutional forms, an agreement on rules and a rule enforcing institution, the protective state, is achieved simultaneously. Building on the notion of an agent of collective action in the role of the state founder or founding agency a possible answer to the question is outlined in section I. However, the suggested solution poses an old problem anew and perhaps more explicitly than usually in the literature.

On the one hand, for the constitutional contract to be agreeable, it would certainly have to constrain the power of the agency. If it did not, everyone, except the members of the agency, would run the risk of being deprived of all benefits from avoiding anarchy or of suffering an even worse fate. On the other hand, if the constitutional rules and rights are not to be violated, the agency must be able to police them and to threaten violators with effective sanctions. It must therefore have coercive powers greater than those available to other members of society. How can constraints on the agency's power then be made credible? Rules intended to constrain the agency may be written into the constitution, but who is in the position to enforce them except the agency itself? How can the members of society at the constitutional stage be convinced that the agency will not, once it is established, use its power to further private interests? This old dilemma, often alluded to (most recently, e.g., in North 1990: 59f.) but apparently not considered becoming relevant, is investigated together with its implications for the contractarian approach in section II. The final section offers some tentative conclusions.

I. The Emergence of a Protective Agency

In his reconsideration of Locke's state of nature scenario Nozick (1974: chap.2) argues that individual vulnerability to deceptive activities of others and the inherent instability of mutual-protection associations are likely to result in the formation of protective agencies, one in each geographically distinct area. Such an agency takes over, he submits, the functions of detection, apprehension, judicial determination of guilt, punishment, and exaction of compensation (a similar assessment is

given by Buchanan 1975: chap.1). Obviously, a protective agency must be properly endowed with the power to protect and to enforce what has been determined. It must be able to threaten or actually use coercive means.

To present the conjecture of an emerging protective agency in a more rigorous form, the point of departure, a state of nature or 'Hobbesian jungle', can be modelled as a prisoner's dilemma game (Wagner and Gwartney 1988; Schmidt-Trenz 1989; Okada and Kliemt 1991). Although, in principle, each member of society would fare better if everyone behaved in a civilized and cooperative manner (strategy y) individual interactions are likely to be characterized by violence and defecting (strategy x) and desirable behavior will therefore not be observed. Let the set $M = \{1,...,n\}$ denote the members of society and each $i \in M$ for convenience be treated here as identical. If n is large and if there are no particular reasons for believing that the members have preferences for interacting with specific persons, state-of-nature interactions can, in first approximation, be broken up in a series of pairwise one-shot interactions with randomly mated agents and without personal recognition of earlier mates.

Hence, any interaction in the series, involving two agents $i \in M$ and $j \in M$, $i \neq j$, is a standard symmetric 2×2 prisoner's dilemma game of the form

$$\Gamma = \{(i,j), (S_i, S_j), (u_i, u_j)\} \tag{1}$$

in which the strategy set is

$$S_i = S_j = \{x, y\} \tag{2}$$

with strategies x and y as just mentioned. The pay-offs, measured in terms of a standard Neumann-Morgenstern utility function, $u_i = u_i(s_i, s_j)$, $s_i \in \{x, y\}$ and $s_j \in \{x, y\}$, satisfy the order relation

$$u_i(x_i y_j) > u_i(y_i y_j) > u_i(x_i x_j) > u_i(y_i x_j) \tag{3}$$

and for u_j, reversing the indices of the strategies in (3), respectively. Since for any agent i the outcome of choosing x_i dominates y_i in utility terms whatever opponent j does, strategy x will be adopted by both parties resulting in the Nash equilibrium pay-off $u_i(x_i x_j)$. Neither $u_i(x_i y_j)$ nor the Pareto superior pay-off $u_i(y_i y_j)$ will be feasible. Any attempt to change this can only be successful if it brings about a corresponding modification of the ordering (3).

A protective agency may indeed achieve this. Imagine someone, motivated to, and capable of, taking the role of an agent of collective action who initiates a voluntary association with the following rules. Every individual in society is free to join the association provided (s)he subscribes to the statutes and pays a membership fee. Statutes prohibit defection in interactions with other members of the association and state that violations will be punished. Membership fees are used to establish and support a task force, the protective agency, supplying effective enforcement, punishment measures, and exacting compensation.[1] The option of joining such an association extends the choice set of the agents. A decision on whether to join *and* paying the fee (decision *f*) or to decline to do so (decision *d*) has to be made before any interactions take place. Conditional on *f* new strategies y' of cooperating or x' of defecting *after* joining and paying the fee are available. Conditional on *d* it is strategies y and x to which the agent is restricted. The corresponding game is

$$\Gamma' = \{(i,j), (S_i', S_j'), (u_i, u_j)\}, \tag{4}$$

where

$$S_i' = S_j' = \{x \text{ or } y | d \text{ and } x' \text{ or } y' | f\} \tag{5}$$

and

$$u_i = u_i(s_i', s_j'),\ s_i' \in S_i',\ s_j' \in S_j' \tag{6}$$

and u_j analogously. Γ' is depicted in *Figure 1*. Whenever two nonmembers of the association meet they are obviously involved in an interaction as in the original game Γ depicted in cells 1-4 in *Figure 1*.

In order for the association to function it must be possible to identify the membership status of an agent sufficiently safely. For simplicity assume that the status is identified with certainty before any interaction takes place. Under this assumptions dominance relations between x' and y' imply a compound strategy for all agents joining the association as follows. Let the set $A = \{1,\ldots,m\}$, $A \subset M$, denote the members of

[1]This association is a special case of an "economic club" (Buchanan 1965; Sandler and Tschirhart 1980) which provides a particular club good, namely penalizing noncooperative behavior within the association. Unlike the literature on clubs which focusses on allocative and efficiency aspects, the present concern is with the question of how such a "club" can be imagined to emerge.

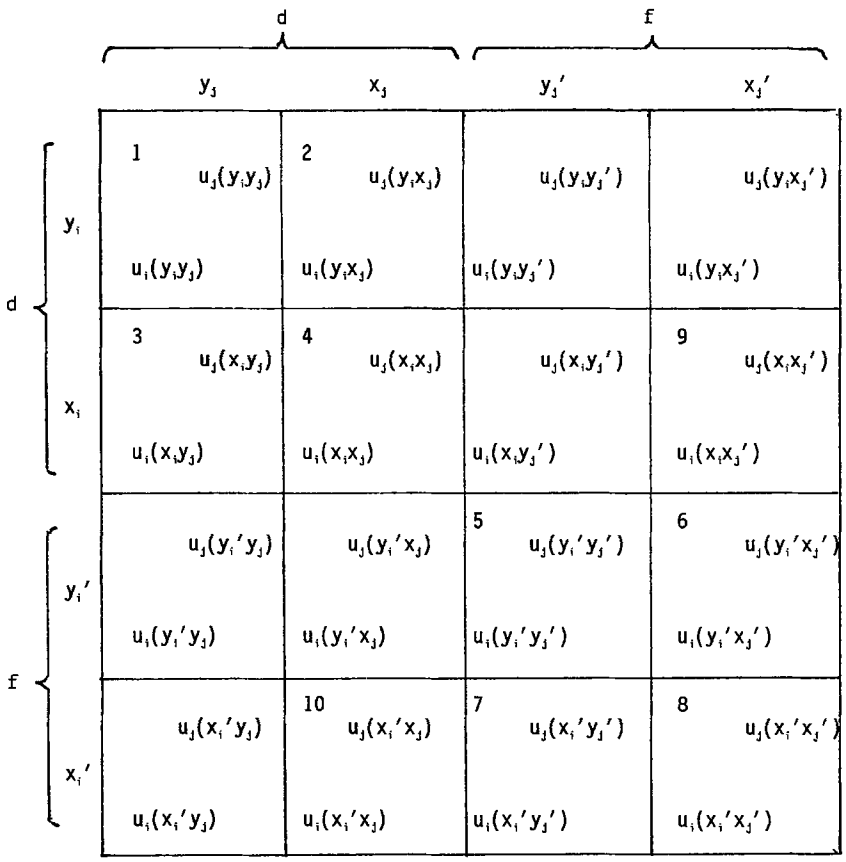

Figure 1. Game Γ'.

the association. If a member i happens to meet a non-member j in an interaction, strategy x_i' dominates y_i' since for obvious reasons the pay-offs satisfy the order relation

$$u_i(x_i'y_j) > u_i(y_i'y_j) > u_i(x_i'x_j) > u_i(y_i'x_j), \tag{7}$$

where the costs of joining the association (membership fees) are already accounted for, i.e. the $u_i(.)$ denote net utility.

Whenever two members i and j of the association meet, however, they are in the position to engage in a modified one-shot interaction with choices and pay-offs as depicted in cells 5-8 in *Figure 1*. Assuming that the association's threat of punishing violations of the statutes and

of exacting compensation is credible, the following pay-off ordering holds:

$$u_i(y_i'y_j') \geq u_i(y_i'x_j') > u_i(x_i'y_j') > u_i(x_i'x_j'). \tag{8}$$

(8) implies that for any member i of the association cooperating dominates defecting irrespective of what member j, with whom interaction takes place, does. Hence, whenever two members meet they will cooperate. Accordingly, upon joining the association the conditional strategy

$$z_i = \{x_i' \text{ if } j \notin A,\ y_i' \text{ if } j \in A\} \tag{9}$$

will be adopted by all members of the association. Because of the membership fee, the following holds:

$$u_i(x_ix_j) = u_i(x_ix_j') > u_i(x_i'x_j) > u_i(y_i'x_j). \tag{10}$$

Since there would be no point in considering an association unless the potential gains from cooperating are sufficiently large, it will be assumed that, with respect to the two best feasible outcomes in (3) and (8),

$$u_i(y_i'y_j') > u_i(x_ix_j) \tag{11}$$

is satisfied.

Given the order relations (3),(7),(8),(10), and (11), the choice between joining and not joining the voluntary association in the first step can be assumed to be associated in the perception of the individual agents with a choice between dominant strategies x and z prompted in the second step. If this consequence is anticipated, the appearance of the voluntary association changes the nature of the strategic interactions. The original prisoners' dilemma game denoted in cells 1-4 is substituted by a coordination (or convention) game represented by cells 4,9,10,5 in *Figure 1*. The coordination game taken by itself has two equilibrium points in pure strategies, $\{y_i'y_j'\}$ and $\{x_ix_j\}$. If, for illustrative purposes, rank order indices are specified in such a way that the order relations (3),(7),(8),(10) and (11) are satisfied and are inserted in place of the pay-offs, the described transition of the game can be demonstrated as in *Figure 2*.

With respect to the individual decision about whether or not to join the association, the crucial point is how likely an interaction with other members of society results in a pay-off $u_i(y_i'y_j')$ or a pay-off $u_i(x_i'x_j)$. Given random mating, this depends on the relative frequency p of members of the association in society, $p = m/n$ (the relative frequency of non-members being $1—p$). Assume for the moment p were known to

Rank order indices of pay-offs in game Γ' assumed for display below

$u_i(x_iy_j) > u_i(y_iy_j) > u_i(y_i'y_j') > u_i(x_ix_j) = u_i(x_ix_j') > u_i(x_i'x_j) > u_i(y_ix_j)$						
8	7	6	5	5	4	2

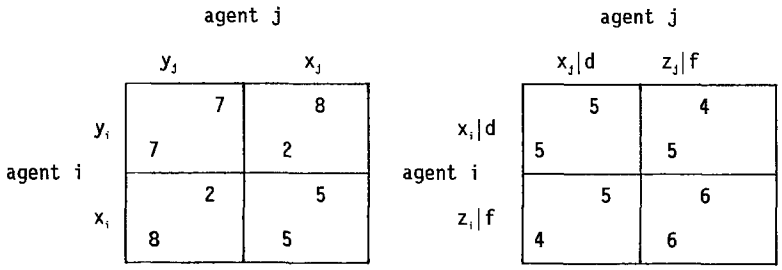

Figure 2. PD-game (left) covered into Coordination game (right).

all agents. The subjectively expected utility $E(U_i')$ accruing to agent i from interactions after joining the association would then depend on p in the following way:

$$E(U_i'|p) = p\, u_i(y_i'y_j') + (1-p)\, u_i(x_i'x_j). \tag{12}$$

The right-hand side of (12) denotes the pay-offs resulting according to the strategy matches in cells 5 and 10 multiplied by the probabilities with which they occur respectively. Alternatively, the subjectively expected utility $E(U_i)$ of interactions after not joining would be

$$E(U_i|p) = (1-p)\, u_i(x_ix_j) + p\, u_i(x_ix_j') = u_i(x_ix_j) \tag{13}$$

by relation (10).

Assuming subjectively expected utility maximization, a necessary and sufficient condition for joining the association is

$$E(U_i'|p) - E(U_i|p) > 0. \tag{14}$$

Solving for p, a threshold value p^* is found such that, if p were known, a decision to join the association would be made once $p > p^*$, but no one would prefer to join otherwise. The critical relative frequency is

$$p^* = \alpha / (\alpha + \beta), \tag{15}$$

where $\alpha = u_i(x_ix_j) - u_i(x_i'x_j)$ measures the loss incurred by paying the fee in vain and $\beta = u_i(y_i'y_j') - u_i(x_ix_j')$ is a measure for the advantage of

cooperation. Because of (8) and (9), $0 < p^* < 1$. Furthermore, the smaller α and/or the larger β, the smaller p^*.

It is rather unlikely, of course, that p is indeed known to the agents. Instead, a subjective expectation $E_i(p) = P_i$ of the unknown p may exist for each $i \in M$. If the decision on joining or not joining the association is made on the basis of P_i, condition (14) may be rewritten as

$$E(U_i'|P_i) - E(U_i|P_i) > 0, \tag{16}$$

and a threshold value P_i^* can be derived in the same way. Each individual is then inclined to join the association once $P_i > P_i^*$. The question is how the subjective expectations are formed. At this point the "agent of collective action" (Olson 1965) may be considered to play a crucial role. Let $M^* = \{1,\ldots,m^*\}$, $m^* > p^*n$ and $M^* \subset M$, be a 'critical mass' of people. If the agent of collective action succeeds in persuading each $i \in M^*$ to expect $P_i > P_i^*$, then the critical mass will indeed be brought together.[2]

This means that once the agent of collective action is able to influence the expectations of a critical mass of members of society her/his foretelling will be self-fulfilling and the supporters' expectations will be confirmed. By the same token, p will exceed p^* so that, after updating individual subjective expectations, everyone else will wish to join the association, and the protective agency will emerge. As can be seen, this is not impossible but may be less easily, and less frequently, achieved than perhaps presumed in the contractarian literature. The transition is certainly not a spontaneous one in the sense of nobody being required to plan and organize it. On the contrary, as already argued by Taylor (1982: 133f.) strong founding figures may be required—chiefs, prophets, state founders who crusade for the public support of their plans and succeed in gaining strongs of supporters.

III. The Agency as a Constitutional Hazard

Provided a critical mass of supporters has been gathered, the voluntary association discussed in the previous section will be unanimously supported by all members of society. It may therefore be considered a device for overcoming anarchy in the spirit of the contractarian

[2]See Witt (1992) for a more detailed discussion and Kuran (1989) for a similar argument focussing on preference falsification rather than expectation formation.

approach. However, although one problem has been resolved in this way, a new one is created. The protective agency required to give credibility to the voluntary association will have to be endowed with significant power so that it can collect fees, sentence, penalize, and exact compensations. In fact, in the previous section it was assumed that enforcement is undiluted. This means that the agency's capacity to threaten and police must go beyond that of any single agent or group of agents. Given the considerable concentration of physical power, how can its use by the agency be kept under control? How can opportunism and perversion by agency personnel in the pursuit of their own private interests be prevented?

Since the voluntary association and its protective agency are often identified with the protective state (Nozick 1974: chap.5; Buchanan 1975; chap.4), the problem may take on the rather serious form of needing to prevent the state administration from usurping the agency's power. Once the coercive means of the protective state have been established those in command of them may well be tempted to divert, at their own discretion, some of the gains from cooperation by abuse of their power. This may happen through various kinds of corruption, through extorting 'protective' payments, or simply through excessive tax charges. The motivation to do so does not necessarily have to be crudely materialistic. There has hardly been a coup d'état in history in which the gains from usurpation have not been claimed to be associated with some "supreme values" in particular if accompanied by a totalitarian revision of the constitution (Bernholz 1991). As is well known, overthrow of constitutional order and usurpation of power in a more or less conspicuous form are historically far from infrequent, not only in 'dark' ages but also in modern times all over the world. Although the armed forces and other empowered branches of the administration in most of the Western democracies have been loyal in recent times there are notable exceptions even here as, for instance, in Turkey, Greece, or Spain where the country suffered no less than 43 coups d'état between 1814 and 1923.

The constitutional hazard has indeed not gone unnoticed (Buchanan 1975: chap.9; Hayek 1979: 109; Wagner and Gwartney 1988). However, at least in the tradition of the "new contractarianism" the discussion usually focuses on the consequences of constitutionally unconstrained majority voting. Important as this may be, it is still a relatively civilized form of hazard as it presupposes a constitutional use of power. What appears much less civilized is the possibility of a brute usurpation of

power by those to whom it has constitutionally been assigned. If this is indeed a significant hazard, and if there are no certain remedies for it, what consequences can it have at the constitutional stage?

Unfortunately, the essence of the contractarian solution of the anarchy dilemma hinges upon this question. If the members of society anticipate a recalcitrant hazard of this sort, it can easily be shown that unanimous agreement for a transition from anarchy to a constitutional state may disappear. In discussing the coming into being of a voluntary association a critical relative frequency p^* of supporters has been determined in the previous section and the conditions have been discussed under which $p > p^*$. In the definition of p^* in equation (15) the variable β measures the advantage of cooperation. If, due to the constitutional hazard, $\beta \rightarrow 0$, i.e. the gain from cooperation accruing to the individual agent as perceived by her/him is going to zero, $p^* \rightarrow 1$. The critical mass M^* of members of society that must be persuaded becomes much larger, making it harder for an agent of collective action to succeed. In the limiting case $\beta = 0$, where all gains are extorted by the 'protective' agency, $p^* = 1$. This means that even a hundred percent support represents an unstable situation which is likely to erode so that it is hard to believe that anarchy can ever be overcome.

Unless it is argued that the constitutional hazard will be disregarded, or discounted, by the members of society, the contractarian approach thus presupposes some kind of remedy to the constitutional hazard inherent in the concentration of physical power in an enforcement agency. Wagner and Gwartney (1988) suggest either designing political institutions and procedures that reduce the likelihood of the abuse of power or installing an "external authority" with the power to enforce constraints on the behavior of the protective agency. Unfortunately, the effects of both these remedies is unclear. The first is a rather platonic notion as the problem is not so much one of design but one of enforcement, given that the only institution in command of unrivalled physical power in society is precisely the protective agency. The second takes this into account but induces an infinite regress—who prevents the external authority from defaulting or colluding? If there is no other remedy, the contractarian solution to anarchy does not seem to escape the constitutional dilemma.

Conclusion

The present discussion started off by reconstructing, in game-theoretic terms, a basic idea of the new contractarians—the hypothetical

achievement of a social contract in an institution free, constitutional setting. The prisoner's dilemma underlying the anarchic state of nature is transformed into a co-ordination (or convention) game involving different strategies by an agent of collective action who crusades for a voluntary 'fair trade' association, a special form of a club. A protective agency provides a club good (enforcement of the agreement among all club members) to all those voluntarily joining the club and funding the agency (paying taxes). As a consequence, all members of society now have the choice of joining *and* cooperating within the 'club' or of not joining and never to cooperate in any interactions. In this game two equilibrium points exist and, because of the agency costs, a critical mass of members must join to make the transition from a state of nature an unanimous choice for all members of society.

Unfortunately, however, with the creation of a protective agency a new problem emerges. Its capacity to threaten and police, which is necessary to make the transition, only works if the agency has sufficient power. If it enjoys a monopoly in the use of coercive means, it may turn out to be difficult to control and to prevent it from usurping the power in the private interest of its personnel. If this usurpation hazard is anticipated by the members of society, there is an obvious constitutional dilemma. Usurpation can deprive the members of the society, except those in the agency, of (almost all) the benefits from overcoming anarchy. Thus, the protective agency required for the credibility of post-constitutional freedom, peace, and cooperation may turn out to be a major threat to precisely these achievements at the post-constitutional level. In historical perspective usurpation is a recalcitrant hazard even though it may not appear to be so in many current modern constitutional states. What distinguishes their performance from that of modern constitutional states with a different record is an empirical question which needs further inquiry. As far as the logic of the argument is concerned, however, the constitutional dilemma appears to challenge the basic idea of the new contractarians: to legitimize the state by giving conditions under which a hypothetical unanimous agreement to a social contract enforced by the protective state could be reached.

REFERENCES

Bernholz, P. (1991) "The Constitution of Totalitarianism." *Journal of Institutional and Theoretical Economics* 147: 435–440.

Buchanan, J. M. (1965) "An Economic Theory of Clubs." *Economica* 32: 1–14.
Buchanan, J. M. (1975) *The Limits of Liberty. Between Anarchy and Leviathan*. Chicago: The University of Chicago Press.
Hayek, F. A. (1979) *Law, Legislation and Liberty*. London: Routledge.
Kuran, T. (1989) "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice* 61: 41–74.
North, D. C. (1990) *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.
Nozick, R. (1974) *Anarchy, State, and Utopia*. New York: Basic Books.
Olson, M. (1965) *The Logic of Collective Action*. Cambridge, Mass.: Harvard University Press.
Okada, A. and Kliemt, H. (1991) "Anarchy and Agreement—A Game Theoretic Analysis of Some Aspects of Contractarianism." In: R.Selten (ed) *Game Equilibirum Models II*, Berlin: Springer: 164–187.
Rawls, J. (1971) *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.
Sandler, T. and Tschirhart, J. T. (1980) "The Economic Theory of Clubs: An Evaluative Survey." *Journal of Economic Literature* 28: 1481–1521.
Schmidt-Trenz, H. J. (1989) "The State of Nature in the Shadow of Contract Formation." *Public Choice* 62: 237–261.
Scott, G. H. (1976) "The New Contractarians." *Journal of Political Economy* 84: 673–590.
Taylor, M. (1982) *Community, Anarchy and Liberty*. Cambridge: Cambridge University Press.
Wagner, R. E. and Gwartney, J. D. (1988) "Public Choice and Constitutional Order." In: J. D.Gwartney and R. E. Wagner (eds) *Public Choice and Constitutional Economics*. London: JAI Press: 29–56.
Witt, U. (1989) "The Evolution of Economic Institutions as a Propagation Process." *Public Choice* 62: 155–172.
Witt, U. (1992) "The Endogenous Public Choice Theorist." *Public Choice* 73: 117–129.