German Economic Review 1(1): 83–110

Co-evolution of Preferences and Information in Simple Games of Trust

Werner Güth
Humboldt University, Berlin

Hartmut Kliemt
Gerhard-Mercator University, GH, Duisburg

Bezalel Peleg
Hebrew University of Jerusalem
and Copenhagen Business School

Abstract. *In standard rational choice modelling decisions are made according to given information and preferences. In the model presented here the 'information technology' of individual decision-makers as well as their preferences evolve in a dynamic process. In this process decisions are made rationally by players who differ in their informational as well as in their preference type. Relative success of alternative decisions feeds back on the type composition of the population which in turn influences rational decision-making. An indirect evolutionary analysis of an elementary yet important basic game of trust shows that under certain parameter constellations the population dynamics of the evolutionary process specify a unique completely mixed rest point. However, as opposed to previous studies of preference formation in the game of trust there is no convergence to but only circumventing the rest point if the informational status of individuals evolves rather than being chosen strategically.*

1. INTRODUCTION

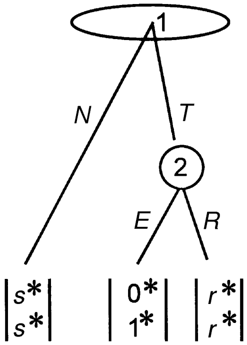
In standard rational choice modelling, decision-makers decide completely opportunistically in view of the anticipated causal consequences of each of their decisions taken separately. In standard models of evolutionary game theory, choices are treated as if being determined by a program or a disposition specifying how to choose in each of a whole class of decision situations. Though models that focus on either of the two extremes have provided valuable insights, in the real world human decision-making is located somewhere between the extremes on which standard models focus. It is influenced by the expected future and by the experienced past. Therefore, in our efforts to model and understand social behavior we should not confine ourselves to isolating one of these factors by abstracting away the other one but rather try to reach the middle ground between the traditional extreme modelling assumptions. We feel that an indirect evolutionary approach is particularly well suited for this task.

Our subsequent discussion of the co-evolution of preferences and information in a simple game of trust will integrate ‘expectational pull’ and ‘adaptive push’ in a single model. In previous indirect evolutionary analyses of trust problems (see Güth and Kliemt, 1995) we focused on the evolution of preferences while assuming that players could strategically decide about the acquisition of costly information. As opposed to that we shall now treat the emergence of information conditions in basic trust situations as endogenous to the evolutionary process. More specifically, in Section 2 we lay out our basic trust game and the background idea of the indirect evolutionary approach. In Section 3 we consider two extreme information conditions and solve the basic games for both. In one of the two polar extremes players who must decide on whether to trust or not to trust are completely uninformed about their co-player’s type. In the other polar case they are completely informed in that regard. In the next section, 4, it is described how players’ preferences to behave trustworthily and their ability to detect that disposition in others may evolve in a co-evolutionary process. Players are either completely informed I-types whose detection technology truthfully reveals their co-player’s type with complete reliability at cost C or they are totally uninformed U-types whose detection technology is costless but so unreliable that they are left completely in the dark about their co-player’s type. Section 5 considers whether under the extreme information conditions associated with different player types there are any population compositions which are dynamically stable along both the preference and the information dimension. Section 6 generalizes the argument to the intermediate case in which a ‘C, μ’ information technology providing merely a stochastic signal about the co-player’s type is available at cost $C > 0$. Section 7 puts our results into perspective and relates them to other approaches as pursued either by ourselves or others.

2. THE TRUST PREDICAMENT

Owing to the fact that human decision-makers are endowed with the distinctly human faculty to make opportunistically rational choices social interaction is soaked with problems of trust. Mutually advantageous forms of cooperation may be closed off since they require that one partner move first without a guarantee that the other(s) will reciprocate. Thus the ‘trust predicament’ emerges.

The essentials of interaction situations that can give rise to problems of trust are represented in Figure 1. The tree of Figure 1 should not be misinterpreted as



$1^* > r^* > s^* > 0^*$

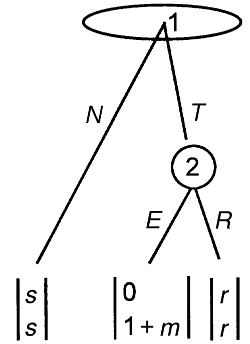
Figure 1

a game in the proper sense of that term. The starred entries at the ends of the branches of the tree are not payoffs representing subjective preferences. They rather represent objective or material rewards as measured for instance in monetary or other units including such possibilities as increases in the number of offspring.

For individuals whose preferences can be represented by a payoff function which in fact coincides with the material payoffs – whose preferences are 'linear in money' – the objective situation of Figure 1 is transformed into a game of trust by simply dropping the 'stars'. In that game, under the conventional background assumption of common knowledge of the tree, rational choice dictates that the first mover choose N since he foresees that the second mover rationally chooses E. Both players end up with s in terms of their subjective evaluations and with s^* in objective terms while they could have realized $r > s$ or $r^* > s^*$ respectively. The Pareto superior result is closed off.

There are several conceivable 'solutions' for problems of trust (see, for fine overviews over social situations involving problems of trust, Fukuyama, 1995; Landa, 1994; Seligman, 1997). The institutions of promise and contract form a case in point. These institutions confer commitment power on individuals (see, on the fundamental role of power conferring rules, the seminal analysis in Hart, 1961). Being endowed with commitment power individuals can choose to modify their expected future payoffs in a preceding strategic move. To put it otherwise, by their preceding (contractual) promise individuals can render themselves 'trustworthy' or – more formally speaking – choose to play a modified subgame.

Institutional solutions to trust problems are what rational-choice theorists would focus on naturally. Other social theorists have traditionally drawn attention to the 'internalization of norms'. Such an internalization gives rise to a modification of preferences as well, however, without a preceding strategic choice of the affected individual. We feel that rational-choice theorists, rather



$1 > r > s > 0$

Figure 2

than rejecting the 'internalization of norms' out of hand as obscure or ad hoc, should better seek ways to integrate it in their models. Besides extrinsic motivation which rests on the subjective perceptions of the objective reward structure of an action situation, intrinsic motivation which derives from other sources like norm internalization can be a central element of decision-making (see, in an economic vein on intrinsic motivation, Frey, 1997).

In Figure 2 a purely subjective motivational component not based directly on objective rewards shows up as the payoff component, *m*. Wheeling in such a motivational factor like *m* would be purely ad hoc if we let it rest with that. But we shall treat *m* as an endogenous parameter of our model. In doing so we distinguish two player types, the trustworthy $\underline{m}$-type for whom in the role of the second mover due to $\underline{m} + 1 < r(\underline{m} < r - 1 < 0)$ the choice of *R* is rational and the untrustworthy $\overline{m}$-type for whom, because of $\overline{m} + 1 > r(\overline{m} > r - 1)$, the choice of *E* is rational. Since all values of *m* that fulfill the same preceding inequalities are behaviorally equivalent it suffices to focus on two types. Accordingly in our discussion of the evolution of the population's type composition each of the two types represents a whole interval of behaviorally equivalent values of *m*.

In previous work we considered how type-dependent differences between the trustworthy and untrustworthy types' decision-making would bring about differential evolutionary success – as measured along the objective dimension of material reward. We studied the evolution of the composition of a population of trustworthy and untrustworthy types under various parameter constellations and modelling assumptions. But we focused exclusively on the evolution of preferences. With respect to information we relied on the assumption that players could choose strategically whether or not to acquire information of reliability μ at some cost *C*. Now we assume that informed I-type players who are endowed with a 'C, μ' technology emerge in the evolutionary process and survive if they are more successful than the uninformed U-type players who do not incur the cost but remain totally uninformed about their co-player's type. Unlike other evolutionary approaches (see for instance the discussion of the evolutionary dynamics of crime in Cressman *et al.*, 1998) ours is still indirect in other regards. Even though the informed players simply emerge, rational strategic choices in the stage games are still explicitly modelled rather than eliminated altogether.

3. BASIC GAMES OF TRUST AND THEIR SOLUTION

3.1. Basic games of trust as played by the uninformed and the informed

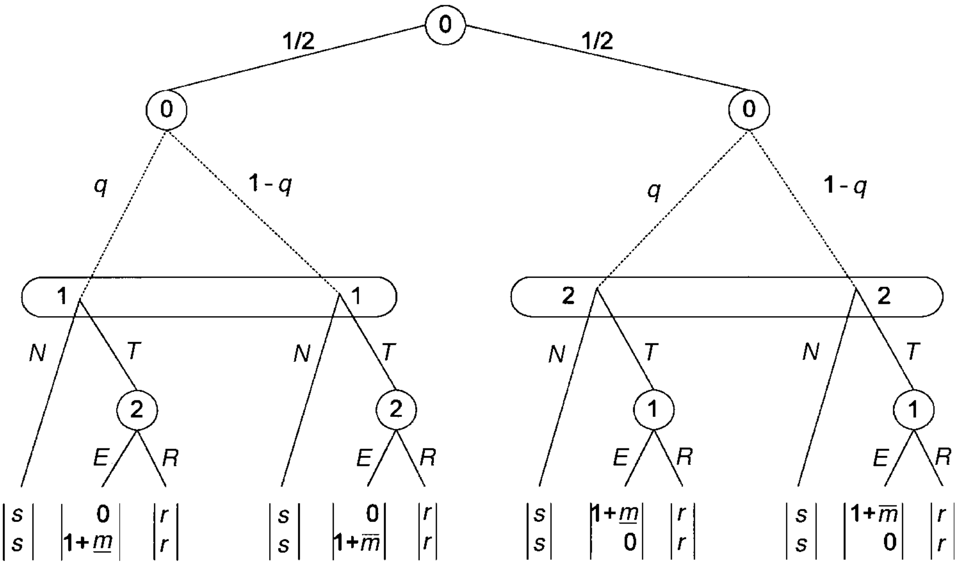
Imagine a game of trust as described in Figure 2. Assume further that the game is played by U-type players who do not know their co-player's preference parameter m or, as we shall also say, they are ignorant of their co-player's 'moral' type. Assignment of moral types can be modelled in the conventional way as the outcome of a fictitious random move in which 'nature' or 'player 0' chooses the types of the players. Since only the trustworthiness of the second mover affects the outcome of the game we can neglect the 'moral' type of the player who ends up in the first-mover role. We need to model only the type assignment for the second-mover role.

In our model, 'nature' or 'player 0' starts by assigning first- and second-mover roles to players 1 and 2, respectively, in a random move with an equal likelihood of 1/2 for either of the two roles. With probability q, corresponding to the share q of trustworthy $\underline{m}$-types in the population, player 0 'chooses' in the previously mentioned fictitious random move that the second mover be a trustworthy $\underline{m}$-type. Afterwards the game of trust of Figure 2 is played according to type and role assignment as fixed in the preceding random moves. The tree shown in Figure 3 emerges.

As indicated above, this tree models the game among U-type players. When assigned the first-mover role their detection technology does not reveal any information on the type of the second mover. However, if there exist I-type individuals then these types in the role of the first mover – due to their perfect detection technology – have complete information on the second mover's type. With an I-type as player 1 and a U-type as player 2 the game is characterized by the graph shown in Figure 4. As for the other assignments of information types: if only player 2 is an informed type her information sets become singletons, while player 1 cannot discriminate. If both players are informed then all information sets are singletons.

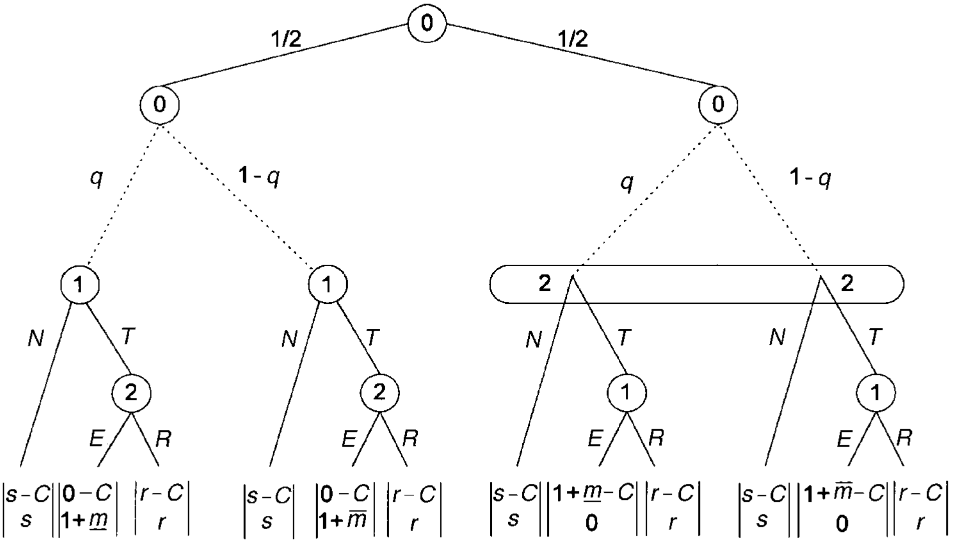
3.2 Solutions of trust games played by different informational and moral types

Assume that the population share of I-types is $p \in [0, 1]$ while that of U-types is $1 - p$. Assume further that the population share q of trustworthy $\underline{m}$-types



$m_i = \underline{m}$ or $m_i = \overline{m}$ for $i = 1, 2$ and $\underline{m} < r - 1 < \overline{m}$, $1 > r > s > 0$

Figure 3 U-type player 1 and U-type player 2



$m_i = \underline{m}$ or $m_i = \overline{m}$ for $i = 1, 2$ and $\underline{m} < r - 1 < \overline{m}$, $1 > r > s > 0, C \geq 0$

Figure 4 I-type player 1 and U-type player 2

and the complementary share $1 - q$ of untrustworthy $\overline{m}$-types is common knowledge among the players. Looking at the game recursively let us start with the behavior in the role of the second mover. Obviously, for behavior in second-mover roles the informational type does not matter. Only the 'moral' type of being trustworthy or untrustworthy is then relevant: in second-mover roles trustworthy $\underline{m}$-types will choose *R* while untrustworthy $\overline{m}$-types choose *E*.

On the other hand, in the first-mover role the moral type of the decision-maker is behaviorally irrelevant. But choosers of different informational types tend to choose differently in first-mover roles. The uninformed U-type, who does not have access to specific type information and therefore must rely on the prior probability q of meeting a trustworthy second mover will choose to trust if $qr + (1 - q)0 > s$ (i.e. $q > s/r$) and not to trust if $qr + (1 - q)0 < s$ (i.e. $q < s/r$) (as shall subsequently become obvious: the degenerate case $q = s/r$ can be neglected for the purposes at hand). The decisions of the fully informed I-type will, however, be determined completely by his knowledge of the second mover's type: if the second mover is a trustworthy $\underline{m}$-type, an I-type first mover, expecting r rather than s, chooses *T*; if the second mover is an untrustworthy $\overline{m}$-type, the I-type first mover, expecting s rather than 0, decides on *N*.

In principle we have to consider four combinations of informational and moral aspects and thus four player types: $\underline{m}$-I-types, $\underline{m}$-U-types, $\overline{m}$-I-types and $\overline{m}$-U-types. However, since in first-mover roles players independent of their own moral type choose according to their informational type whereas in second-mover roles moral type is all that matters is the moral and informational dimensions of the problem are 'separable'. It suffices in the subsequent discussion of the evolutionary dynamics of the population composition to consider the relative survival prospects of U- vs I-types independently of their own moral type while the relative success of the $\underline{m}$- vs $\overline{m}$-types does not depend on their informational type.

4. EVOLUTIONARY DYNAMICS

Let us now investigate how p and q evolve if the preceding games are played among randomly paired players chosen from an infinite population of individuals of the four 'moral-cum-informational' types. Our basic assumption is that q increases (decreases) when the $\underline{m}$-types are more (less) successful than $\overline{m}$-types in basic games. Similarly, p is assumed to increase (decrease) whenever I-types are more (less) successful than U-types.

Since we assume that players are randomly drawn from a population containing infinitely many individuals of all types the average expected success of different types will have 0 variance, i.e. each type's average success is not stochastic while that of the individuals belonging to the different types may well be. In view of this we can directly relate the decrease and increase in population shares p and q to expected payoffs and these in turn to their objective ('starred') counterparts – the 'currency' in which relative success in evolutionary competition is ultimately determined.

Now, starting with relative success of informational types, consider the expectations of an I-type as opposed to those of a U-type. For the sake of specificity assume, initially, that the situation sketched in Figure 4 obtains. Player 1 is an I-type and player 2 a U-type. Player 1 has incurred a sunk cost of C. Whether or not this makes him better off than a U-type player 1 depends on the value of the information to which he gains access. So let us look at the value of the information.

With a probability of $\frac{1}{2}$ player 1 expects to play as a first and with equal likelihood as a second mover. Before matching and before role assignment player 1's expectations for the contingency of becoming the first mover depend solely on the population composition parameter q. This parameter describes how likely it is that he be paired with a trustworthy co-player whom he then would recognize since, as an I-type first-mover, player 1 can perfectly discriminate between trustworthy and untrustworthy second movers. Thus before matching and role assignment I-type player 1 expects $\frac{1}{2}[qr + (1-q)s]$ from the first-mover role. As compared to that a U-type player 1 would expect $\frac{1}{2}[qr + (1-q)0]$, for $q > s/r$, and $\frac{1}{2}s$, for $q < s/r$.

When assigned the second-mover role I-type player 1's information about first-mover player 2's type is strategically irrelevant. The I-type, like the U-type second mover, plays solely according to his own moral type. In the case described in Figure 4 both types of player 1 are dealing with an uninformed player 2 in the first-mover role. Behavior of such a U-type player depends solely on the relation between q and s/r. Thus, under the information conditions depicted in Figure 4, I-type and U-type players of identical moral type have identical expectations in second-mover roles.

It is obvious that the preceding line of argument applies to all combinations of types assigned to players 1 and 2: I-type and U-type players of identical moral type have identical expectations in second-mover roles as determined by their co-player's informational type (and, of course, independently of the first-mover co-player's moral type).

If the identical expectations of I-types and U-types in second-mover roles are W then before role assignment and matching the expectation of both informationally different but morally identical types of player 1 are $\frac{1}{2}W$. The same term $\frac{1}{2}W$ forms part of the expectation $R_{\mathrm{I}}(q)$ of the informed I-type and of the expectation $R_{\mathrm{U}}(q)$ of the uninformed U-type. Therefore the payoff in second-mover roles cannot differentiate between morally identical types that differ only along the informational dimension. Payoff in second-mover roles can therefore be neglected in the determination of relative evolutionary success of morally identical alternative informational types.

Taking into account this observation we can focus solely on differential success of alternative informational types in first-mover roles. Moreover, since we are dealing with an infinite population in which expectations directly translate into relative success of different types we get – after multiplying by 2 – the following simplified success functions for I-types and U-types:

$$R_{\mathrm{I}}(q) = qr + (1-q)s - 2C \tag{4.1}$$

$$R_{\mathrm{U}}(q) = \begin{cases} s & \text{for } q < s/r \\ qr & \text{for } q > s/r \end{cases} \tag{4.2}$$

For $R_{\mathrm{I}}(q) > R_{\mathrm{U}}(q)$ to be fulfilled it is necessary that

$$2C < \frac{s}{r}(r-s) \tag{4.3}$$

and

$$\frac{2C}{r-s} < q < \frac{s-2C}{s} \tag{4.4}$$

apply. The left-hand side of (4.4) emerges for $s/r > q$ and the right-hand side for $s/r < q$. If the interval for q characterized by (4.4) is non-empty, this in turn implies that (4.3) obtains as well. Moreover, if (4.4) holds good and thus the necessary condition, (4.3), for a non-empty interval applies as well, then $q = s/r$ will always be an element of that non-empty interval.

The negation of (4.3) is $2C \geq (s/r)(r-s)$. If $2C$ is strictly greater than the right-hand side of the inequality there are no populations of trustworthy and untrustworthy types such that the informed would do better than the uninformed type. Due to the high 'fixed cost' of being an informed type, I-types are eventually driven out of the population. In the degenerate case in which (4.3) becomes an equality with an empty interval (4.4), the population share p of informed individuals will almost always decline too.

Up to now we relied on the sign of $R_{\mathrm{I}}(q) - R_{\mathrm{U}}(q)$ as indicating the direction of change. We have for all $t \geq 0$:

$$R_{\mathrm{I}}(q_t) - R_{\mathrm{U}}(q_t) = \begin{cases} (r-s)q_t - 2C & \text{for } q_t < s/r \\ s - 2C - sq_t & \text{for } q_t > s/r \end{cases} \tag{4.5}$$

Clearly, under any plausible specification of the dynamic process, as long as $R_{\mathrm{I}}(q_t) - R_{\mathrm{U}}(q_t) > 0, p_t$ should increase through time.

To say anything about the dynamic process itself we need to specify it. So let us postulate here without further justification that the dynamic process can be described by a linear relationship:

$$\dot{p}_t = k[R_{\mathrm{I}}(q_t) - R_{\mathrm{U}}(q_t)] \tag{4.6}$$

where k is a positive constant. (In a somewhat different but related context p_t may be interpreted as a mixed strategy whose dynamics is justified in a learning context as for instance in Arthur, 1993, and Posch, 1997, who refer to a Polyatype urn model while other stochastic dynamics like those discussed in Hofbauer and Schlag, 1998, would be inconsistent with the rational-choice assumptions introduced in Section 3.2 above.)

Turning to the dynamics of q let us assume initially that (4.3) obtains and that there is a permanent influx (mutation) of I-types such that there are always some informed individuals around. Thus, even if $p_t \to 0$ for $t \to \infty$, the population share p_t of informed I-types is always positive.

In second-mover roles differential success of moral types depends on the population share p of informed types in first-mover roles. For an $\underline{m}$-type second mover we therefore get:

$$R_{\underline{m}}(p) = \begin{cases} pr + (1-p)s & \text{for } q < s/r \\ r & \text{for } q > s/r \end{cases} \tag{4.7}$$

and for an $\overline{m}$-type

$$R_{\overline{m}} = \begin{cases} s & \text{for } q < s/r \\ ps + (1-p)1 & \text{for } q > s/r \end{cases} \tag{4.8}$$

$R_{\underline{m}}(p) - R_{\overline{m}}(p) > 0$ is fulfilled for $q < s/r$ since, by assumption, $p > 0$ and $r > s$. Therefore, for $q < s/r$ and $p > 0$ the share q of $\underline{m}$-types increases while for $q > s/r$:

$$p > \frac{1-r}{1-s} \tag{4.9}$$

is required for that to be the case. Thus an increase of q is to be expected unless $q > s/r$ and (4.9) is reversed. In the latter case, $R_{\underline{m}}(p) - R_{\overline{m}}(p) < 0$ and q will decrease.

Again, in view of the difference

$$R_{\underline{m}}(p) - R_{\overline{m}}(p) = \begin{cases} (r-s)p_t & \text{for } q < s/r \\ (1-s)p_t - (1-r) & \text{for } q > s/r \end{cases} \tag{4.10}$$

we specify the dynamic process by postulating the linear relationship

$$\dot{q}_t = h[R_{\underline{m}}(p_t) - R_{\overline{m}}(p_t)] \tag{4.11}$$

where h is a positive constant.

If the reverse of relation (4.3) applies we know from the preceding argument about the dynamics of p that the informed I-types will tend to be driven out of the population entirely. However, if one assumes – as we did above and shall do after the following brief digression – that there is a permanent influx of I-types such that $p_t > 0$ then $p_t = 0$ is merely a limit result. Still, it may be noted that $p_t = 0$ corresponds to a generic case, if the type-revealing signal that is received in first-mover roles becomes stochastic. Therefore it seems appropriate to look at the limiting constellation somewhat more carefully rather than to dismiss it out of hand.

Suspending the assumption $p_t > 0$ let us for the time being and for the sake of simplicity suppose that I-types have been driven out completely and therefore all players are uninformed U-types. Though all are U-types some may be trustworthy and others untrustworthy. However, since all are uninformed the game of Figure 3 is played throughout. Now, it is obvious that in the game as displayed in Figure 3, untrustworthy second movers – regardless of their informational characteristics in first-mover roles and the costs they would incur then – have an advantage over trustworthy ones if they find a trusting first mover. In first-mover roles uninformed first movers, regardless of their own moral type, will rationally choose to trust as long as $q > s/r$. Thus, in that realm in a population of uninformed U-types the population share q of the trustworthy will decline. On the other hand, since in case $q < s/r$ first movers should rationally choose N throughout, nothing could differentiate between types leaving the population composition unaltered. Therefore all (p, q) with $p = 0$ and $q < s/r$ would be rest points, though with degenerate attraction sets.

Assuming that decision-makers, even though they are rational, are prone to commit occasional mistakes uninformed first movers would occasionally mistakenly choose to trust even if $q < s/r$. Then the untrustworthy could have an advantage over the trustworthy types beyond the threshold of $q = s/r$ and the population share q of trustworthy individuals would tend to decline over the whole interval (0, 1], though the decline would be slowed down for $q < s/r$. Figures 5 and 6 sum up the preceding discussion of the dynamics of p and q.

5. STABILITY OF *p, q* CONSTELLATIONS

A rest point (p^*, q^*) of the p, q dynamics is characterized by vanishing rates of change $\dot{p}_t$ and $\dot{q}_t$:

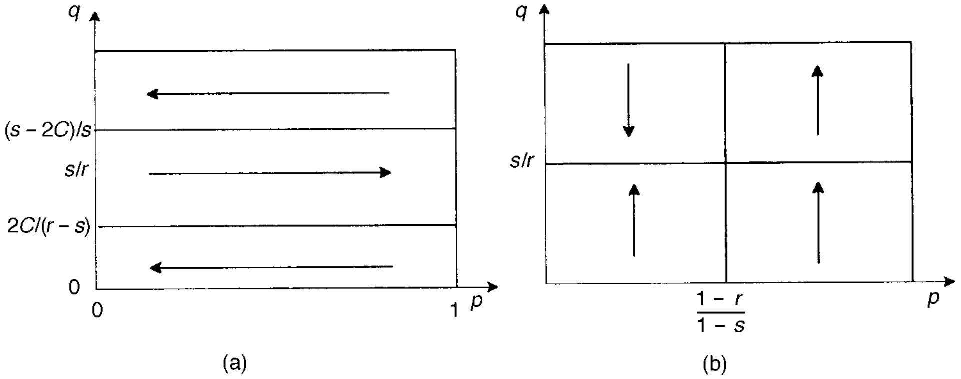


Figure 5a, b The direction in which p and q change for $2C < (s/r)(r - s)$

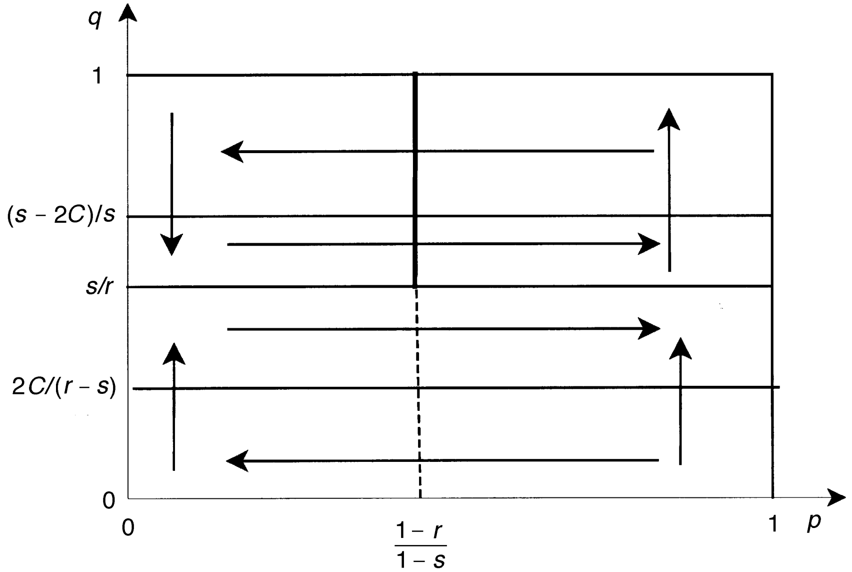


Figure 6 The p, q dynamics for $2C < (s/r)(r - s)$

$$(\dot{p}_t, \dot{q}_t) = (k[R_{\mathrm{I}}(q^*) - R_{\mathrm{U}}(q^*)], h[R_{\underline{m}}(p^*) - R_{\overline{m}}(p^*)]) = (0, 0) \tag{5.1}$$

Since $p_t > 0$ and $q_t < s/r$ imply $\dot{q}_t > 0$ there are no rest points

$$(p, q) \in [0, 1]^2 \quad \text{with } q < s/r \text{ and } p > 0 \tag{5.2}$$

Therefore in the range $q < s/r$ only the points from the line segment $[p = 0, q \leq s/r)$ or – in case of trembles (on how 'general trembles' may affect results, see Gale *et al.*, 1995) – possibly the point $(p, q) = (0, 0)$ can be rest points (with degenerate attraction sets). In the degenerate case $q = s/r$ the first moving player is indifferent between N and T. There is no unique solution play for the first moving player. Since changes in the population composition depend on play and since the first-mover's play cannot be predicted, $(\dot{p}_t, \dot{q}_t)$ is not determined. Hardly anything can be said about this case.

Other p, q constellations can be at rest only if they are an element of

$$E := \{(p, q) \in [0, 1]^2 | q > s/r\} \tag{5.3}$$

Moreover, if $(p, q) \in E$ and thus $q > s/r$, then $p > (1 - r)/(1 - s)$ implies $R_{\underline{m}}(p) - R_{\overline{m}}(p) > 0$ and $p < (1 - r)/(1 - s)$ implies $R_{\underline{m}}(p) - R_{\overline{m}}(p) < 0$ (see Figures 5 and 6). From this consideration it is already obvious that rest points in E must fulfill $p = (1 - r)/(1 - s)$. Analogous arguments show that it is necessary for a rest point that $q = (s - 2C)/s$ be fulfilled, too. Thus there can only be one point that fulfills both necessary conditions. Since together the two conditions imply vanishing rates of change along both dimensions the unique rest point in E is

$$(p^*, q^*) = \left(\frac{1-r}{1-s}, \frac{s-2C}{s}\right) \tag{5.4}$$

In the rest point (p^*, q^*) the rates of change $\dot{p}_t, \dot{q}_t$ vanish. Intuitively speaking there is no 'push away' from (p^*, q^*) since for sufficiently small perturbations the dynamic process will remain arbitrarily close to that state. More formally, (p^*, q^*) is stable in the sense that for every $\epsilon > 0$ and the corresponding neighborhood under the Euclidean norm $d(\cdots)$ there is a $\delta > 0$ and a corresponding neighborhood of (p^*, q^*) such that for a 'starting point' (p_0, q_0) with $d((p^*, q^*), (p_0, q_0)) < \delta$ we have $d((p^*, q^*), (p_t, q_t)) < \epsilon$ for all $t \geq 0$.

Though stable, the point (p^*, q^*) is, however, not (locally) asymptotically stable. To be (locally) asymptotically stable there must not only be no 'push away' from (p^*, q^*) but rather some 'pull towards' it. More formally, (p^*, q^*) is *not* (locally) asymptotically stable, since there exists no $\epsilon > 0$ such that for all starting points (p_0, q_0) with $d((p^*, q^*), (p_0, q_0)) < \epsilon$ we have $\lim_{t\to\infty}(p_t, q_t) = (p^*, q^*)$. These claims are demonstrated in the subsequent discussion of the time path of (p_t, q_t), $t \geq 0$, for all initial $(p_0, q_0) \in E$.

For $(p_t, q_t) \in E$ we have $q_t > s/r$. Inserting (4.5) into (4.6) and (4.10) into (4.11) yields

$$\dot{p}_t = k[s - 2C - sq_t] \quad k > 0 \tag{5.5}$$

$$\dot{q}_t = h[(1-s)p_t - (1-r)] \quad h > 0 \tag{5.6}$$

Solving (5.5) for q_t and taking the derivative with respect to time yields

$$\dot{q}_t = -\frac{1}{ks}\ddot{p}_t \tag{5.7}$$

and, using (5.7) on (5.6):

$$\ddot{p}_t + khs(1-s)p_t = khs(1-r) \tag{5.8}$$

A constant solution to the inhomogeneous differential equation (5.8) is

$$p_t = c \quad \text{where } c = \frac{1-r}{1-s} \tag{5.9}$$

For the homogeneous equation

$$\ddot{p}_t b^2 p_t = 0 \quad \text{with } b^2 = khs(1-s) \tag{5.10}$$

two independent solutions are given by

$$p_t = \sin(tb) \quad \text{and} \quad p_t = \cos(tb) \tag{5.11}$$

All solutions can be expressed as

$$p_t = c + \lambda_1 \sin(tb) + \lambda_2 \cos(tb) \quad \text{for some } \lambda_1, \lambda_2 \in R \tag{5.12}$$

Taking the derivative

$$\dot{p}_t = \lambda_1 b \cos(tb) - \lambda_2 b \sin(tb) \tag{5.13}$$

using it on (5.5) and solving for q_t we have

$$q_t = \frac{s - 2C}{s} + \frac{b}{ks}[\lambda_2 \sin(tb) - \lambda_1 \cos(tb)] \tag{5.14}$$

From (5.12) and (5.14) it is obvious that the p_t, q_t dynamics are periodic since sin and cos are periodic. Regardless of the two positive factors k and h, any initial p_t, q_t constellation will be revisited indefinitely. Moreover, from any initial $(p_t, q_t) \neq (p^*, q^*)$ with $q_t > s/r$ the dynamic process will not be attracted to the unique rest point $(p^*, q^*) = [(1-r)/(1-s), (s-2C)/s]$ in E. The attraction set of (p^*, q^*) is $\{(p^*, q^*)\}$.

In view of general mathematical insights into the structure of (time) continuous processes in R^2 this result is hardly surprising. A process starting at some $(p, q) \in E$ must either converge to some (p^*, q^*) or be periodic. In our case periodicity ensues because linearity and, for that matter, 'separability' prevent convergence. (For related illustrations of the same basic facts in a direct social evolutionary setting see Friedman, 1991; Cressman *et al.*, 1998; and Posch, 1997.)

The preceding considerations rule out any stronger statements about attractors of the dynamic process. We can, however, characterize the periodic process somewhat further. In doing so let us first define two other auxiliary constants

$$d := \frac{ks}{b} \quad \text{and} \quad e := k\frac{s - 2C}{b}$$

Using these constants we can rewrite (5.12) and (5.14) as

$$p_t - c = \lambda_1 \sin(tb) + \lambda_2 \cos(tb) \tag{5.12$'$}$$

$$dq_t - e = \lambda_2 \sin(tb) - \lambda_1 \cos(tb) \tag{5.14$'$}$$

After multiplying the first equation by λ_2, the second by λ_1 and subtracting the results we get

$$\cos(bt) = \frac{\lambda_2(p_t - c) - \lambda_1(dq_t - e)}{\lambda_1^2 + \lambda_2^2} \tag{5.15}$$

whereas multiplying the first equation by λ_1, the second by λ_2 and adding the results yields

$$\sin(bt) = \frac{\lambda_1(p_t - c) + \lambda_2(dq_t - e)}{\lambda_1^2 + \lambda_2^2} \tag{5.16}$$

Using $\cos^2(bt) + \sin^2(bt) = 1$ we get

$$[\lambda_2(p_t - c) - \lambda_1(dq_t - e)]^2 + [\lambda_1(p_t - c) + \lambda_2(dq_t - e)]^2 = [\lambda_1^2 + \lambda_2^2]^2$$

and from this

$$(p_t - c)^2 + (dq_t - e)^2 = \lambda_1^2 + \lambda_2^2 \tag{5.17}$$

The p_t, q_t process cycles on an ellipse around the unique rest point in E, namely $(p^*, q^*) = [(1-r)/(1-s), (s-2C)/s] = (c, e/d)$. The time span for a complete cycle, Δt, can be calculated by considering $p_t = p_{t+\Delta t}$ in equation (5.12′); that is

$$\lambda_1 \sin(tb) + \lambda_2 \cos(tb) = \lambda_1 \sin([t + \Delta t]b) + \lambda_2 \cos([t + \Delta t]b) \tag{5.18}$$

Owing to the periodicity 2π of sin as well as of cos we can conclude that $2\pi = \Delta tb$ and thus

$$\Delta t = 2\pi/b = \frac{2\pi}{\sqrt{khs(1-s)}} \tag{5.19}$$

The two parameters k and h allow for different volatilities of the adaptive processes along the informational or p-dimension and the moral or q-dimension of the population composition. The length of the 'cycle' of the dynamic process in which the population composition changes through time is determined as a function of the product $khs(1-s)$: the larger the factor kh the shorter the cycle. Analogously, the larger the factor $s(1-s)$, i.e. the closer s is to $\frac{1}{2}$, the shorter the cycle.

Finally, for all k, h and any starting point $(p_0, q_0) \in E$ the coefficients are uniquely determined by equations (5.12′) and (5.14′) as

$$p_0 - c = \lambda_2 \tag{5.20}$$

$$dq_0 - e = -\lambda_1 \tag{5.21}$$

where c is a function of the exogenous parameters r, s, while b and d are functions of k, h and again the exogenous parameter s. For all k, h and all $(p_0, q_0) \in E$ the time trajectories of the p_t, q_t process can be determined. As (5.17) shows, this process stays on the same graph while its cycle length depends on kh and $s(1-s)$. For all $t > 0$ we have:

$$p_t - c = (e - dq_0)\sin(bt) + (p_0 - c)\cos(bt) \tag{5.12″}$$

$$dq_t - e = (p_0 - c)\sin(bt) + (dq_0 - e)\cos(bt) \tag{5.14″}$$

The results reached thus far are quite robust. As is shown in the mathematical appendix the linear adjustment process specified in equations (4.6) and (4.11) can be generalized and it can also be secured that an adjustment process starting in the interior of E will remain there.

6. IMPERFECT TYPE DETECTION

The preceding discussion focused on polar extremes. Either U-types were assigned first-mover roles and a game with purely private information on the second mover's moral type was played, or I-types playing in the first-mover role could perfectly discriminate between the trustworthy and untrustworthy second movers. Obviously the middle ground between these extremes is of great interest. In situations involving problems of trust the information status of real-world decision-makers more often than not seems neither to correspond to that of fully informed I-types nor to that of completely ignorant U-types. It seems quite realistic that real-world decision-makers command what we shall call a 'C, μ' technology. This information 'technology' – using the term 'technology' in its widest sense including long acquaintance, detective services, other individuals' looks etc. – provides some information of the second mover's type. More specifically, partially informed I′-type individuals have access to a stochastic signal which reveals the co-player's type with probability $\mu > \frac{1}{2}$. Gaining access they incur, however, some sunk cost $C > 0$.

If the second mover is a trustworthy $\underline{m}$-type then an imperfectly informed first mover receives with probabilty $\underline{\mu}$ the signal '$\underline{m}$'. With probability $(1 - \underline{\mu})$ the signal received at cost $C > 0$ from an $\underline{m}$-type is '$\overline{m}$'. If the second mover is an $\overline{m}$-type then with probability $\bar{\mu}$ the true type is revealed by the signal '$\overline{m}$' while with complementary probability $1 - \bar{\mu}$ the signal '$\underline{m}$' misleadingly indicates a trustworthy type even though it is received from an untrustworthy $\overline{m}$-type.

We require

$$\tfrac{1}{2} < \underline{\mu}, \bar{\mu} \leq 1 \tag{6.1}$$

Knowing the population share q of trustworthy $\underline{m}$-types in the population, all players should rationally endorse the prior probability q of encountering a trustworthy second mover. Beyond this, U-type players do not command information about their second-moving co-player's moral type. But (incompletely) informed I′-type players in first-mover roles receive a type signal indicating with probability $\mu = \underline{\mu}$, $\bar{\mu}$ their second-moving co-player's true type. Obviously the probability of receiving the signal '$\underline{m}$' is $q\underline{\mu} + (1 - q)(1 - \bar{\mu})$ so that the posterior probability for a trustworthy $\underline{m}$-type individual in the second-mover role after receiving the signal '$\underline{m}$' is $q\underline{\mu}/[q\underline{\mu} + (1 - q)(1 - \bar{\mu})]$. Analogously, the probability of receiving the signal '$\overline{m}$' is $q(1 - \underline{\mu}) + (1 - q)\bar{\mu}$. After receiving the signal '$\overline{m}$' the posterior probability for an untrustworthy $\overline{m}$-type in the second-mover role is $(1-q)\bar{\mu}/[q(1-\underline{\mu})+(1-q)\bar{\mu}]$ while after that signal the posterior probability for a trustworthy $\underline{m}$-type is $q(1-\underline{\mu})/[q(1-\underline{\mu})+(1-q)\bar{\mu}]$.

After receiving signal '$\underline{m}$' an imperfectly informed I′-type should rationally prefer to trust, *T*, over showing no trust, *N*, in the role of the first mover if

$$\frac{q\underline{\mu}}{q\underline{\mu}+(1-q)(1-\bar{\mu})}r > s \tag{6.2}$$

while after receiving the signal '$\overline{m}$' he should rationally prefer *N* over *T* if

$$\frac{q(1-\underline{\mu})}{q(1-\underline{\mu})+(1-q)\bar{\mu}}r < s \tag{6.3}$$

For convenience set

$$R := \frac{q\underline{\mu}}{q\underline{\mu}+(1-q)(1-\bar{\mu})} \quad L := \frac{q(1-\underline{\mu})}{q(1-\underline{\mu})+(1-q)\bar{\mu}} \tag{6.4}$$

Observe that $1 > q > 0$ and $\frac{1}{2} < \underline{\mu}, \bar{\mu} < 1$ imply

$$1 > R > L > 0 \tag{6.5}$$

I′-type players will follow the signal '$\underline{m}$' by choosing *T* only if $R > s/r$, and they will follow the signal '$\overline{m}$' by choosing *N* only if $L < s/r$. Thus I′-type players will follow both signals only if

$$\frac{s}{r} \in (L, R) \tag{6.6}$$

Now, note that the relation $R > s/r$ is equivalent to

$$q > \frac{s(1-\bar{\mu})}{\underline{\mu}(r-s)+(1-\bar{\mu})s} \tag{6.2′}$$

while relation $L < s/r$ is equivalent to

$$q < \frac{s\bar{\mu}}{(1-\underline{\mu})(r-s)+\bar{\mu}s} \tag{6.3′}$$

Set

$$J := \frac{s(1-\bar{\mu})}{\underline{\mu}(r-s)+(1-\bar{\mu})s} \quad \text{and} \quad H := \frac{s\bar{\mu}}{(1-\underline{\mu})(r-s)+\bar{\mu}s} \tag{6.7}$$

The assumption $\frac{1}{2} < \underline{\mu}, \bar{\mu}$, implies

$$0 < J < H < 1 \tag{6.5′}$$

Observe also that $s/r < R \Leftrightarrow q > J$ and $L < s/r \Leftrightarrow q < H$ and thus

$$\frac{s}{r} \in (L,R) \Leftrightarrow q \in (J,H) \tag{6.8}$$

Only for (6.8) can type differentiation take place. For, I′-type players should be expected to behave exactly like U-type players whenever $s/r \notin (L,R)$ – i.e., $s/r \geq R$ or $s/r \leq L$ – and thus also $q \notin (J,H)$ – i.e., $q \leq J$ or $H \leq q$. To see this, assume first that the I′-type players ignore the signal '$\underline{m}$' because of $R \leq s/r$. Note, $(R > s/r \Leftrightarrow q > J) \Leftrightarrow (q \leq J \Leftrightarrow R \leq s/r)$. Moreover $\frac{1}{2} < \underline{\mu}, \bar{\mu}$, implies $(r-s)(1-\bar{\mu}) < \underline{\mu}(r-s)$ and $(r-s)(1-\bar{\mu}) < \underline{\mu}(r-s) \Leftrightarrow J < s/r$. Therefore for $R \leq s/r$ we get $s/r > J \geq q$. Because of $s/r > q$ all U-type players should rationally choose not to trust in the first-mover role while the I′-types should not let the signal '$\underline{m}$' induce them to deviate from the strategy suggested by $q < s/r$. Moreover, since $R > L$ the I′-types should follow the signal '$\overline{m}$' and thus in that contingency should certainly behave like the U-types, too. In sum, nothing can differentiate between I′-types and U-types if $s/r \geq R$.

Turning to the other conceivable violation of (6.8), assume second, $s/r \leq L$. Note $s/r \leq L \Leftrightarrow H \leq q$. Moreover, $\frac{1}{2} < \underline{\mu}, \bar{\mu}$ implies

$$q \geq H \Leftrightarrow q \geq \frac{s}{\frac{(1-\underline{\mu})}{\bar{\mu}}(r-s)+s}$$

The latter implies $q > [s/(r-s+s)] = s/r$. Therefore all U-type players should choose T, if $s/r \leq L$. Note that $s/r \leq L$ is the negation of (6.3). It is therefore not the case that the I′-type should choose N over T after receiving an '$\overline{m}$' signal. He should rather ignore that signal and decide on T. Should the I′-type receive an '$\underline{m}$' signal he should certainly choose to trust if $s/r \leq L < R$. In sum, if $s/r \leq L$ nothing can differentiate between U-types and I′-types.

The upshot of the preceding discussion is that for $q \notin (J,H)$ – or, for that matter, $s/r \notin (L,R)$ – nothing can differentiate between uninformed U-types and (incompletely) informed I′-types. If (6.8) is not fulfilled, all types behave like U-types. Therefore the case $s/r \notin (L,R)[q \notin (J,H)]$ corresponds to that of $p = 0$ for which we have shown elsewhere that – at least in the presence of occasional mistakes – there is a single universal attractor for q, namely $q = 0$ (see Güth and Kliemt, 1994). For $q_t > H$, of course, the dynamic process will eventually yield $q_t \in (J,H)$ whereas for $q_t < J$ the range will never be reached. Since $q_t < J$ implies $R_{\mathrm{U}}(q) > R_{\mathrm{I}'}(q)$ and thus a decrease of p, now the line segment $[p = 0, J)$ or – in case of occasional mistakes – the rest point $(p,q) = (0,0)$ have a generic attraction set, namely all (p,q), with $q < J$. The points $(p,q) = (0,q)$ with $0 < q < s/r$ remain rest points with degenerate attraction sets. In our remaining discussion of the case of incompletely informed I′-type players we focus on $q_t \in (J,H)$.

Studying the dynamics of p and q we must again consider the relative reproductive success of I′-types as compared to U-types and of $\underline{m}$-types as compared to $\overline{m}$-types. The condition $R_{\mathrm{I}'}(q) - R_{\mathrm{U}}(q) > 0$ now reads:

$$\{q[\underline{\mu}r + (1-\underline{\mu})s] + (1-q)[\bar{\mu}s + (1-\bar{\mu})0] - 2C\} - s > 0 \quad \text{if } q < s/r \tag{6.9}$$

$$q > \frac{s(1-\bar{\mu}) + 2C}{\underline{\mu}(r-s) + (1-\bar{\mu})s} \tag{6.9$'$}$$

Similarly,

$$\{q[\underline{\mu}r + (1-\underline{\mu})s] + (1-q)[\bar{\mu}s + (1-\bar{\mu})0] - 2C\} - qr > 0 \quad \text{if } q > s/r \tag{6.10}$$

$$q > \frac{s\bar{\mu} - 2C}{(1-\underline{\mu})(r-s) + \bar{\mu}s} \tag{6.10$'$}$$

Again for notational convenience

$$U := \frac{s(1-\bar{\mu}) + 2C}{\underline{\mu}(r-s) + (1-\bar{\mu})s} \quad \text{and} \quad O := \frac{s\bar{\mu} - 2C}{(1-\underline{\mu})(r-s) + \bar{\mu}s} \tag{6.11}$$

For $q \in (U, O)$ the population share p of I$'$-types should be expected to increase whereas a decrease must be expected for $q \notin [U, O]$. Owing to $C > 0$ we get

$$H > O > \frac{s}{r} > U > J \tag{6.12}$$

as long as

$$2C < \frac{s}{r}(r-s)(\underline{\mu} + \bar{\mu} - 1) \tag{6.13}$$

This being said about the informational dimension let us now turn to the moral dimension. As before, we rely on $p > 0$ when studying the case $q < s/r$.

Recall that moral type matters only in second-mover roles. At the same time differential success of moral types depends on the population share p of informed types in first-mover roles. Thus we get for an $\underline{m}$-type second mover

$$R_{\underline{m}}(p) = \begin{cases} p(\underline{\mu}r + (1-\underline{\mu})s) + (1-p)s & \text{for } q < s/r \\ p(\underline{\mu}r + (1-\underline{\mu})s) + (1-p)r & \text{for } q > s/r \end{cases} \tag{6.14}$$

and for an $\overline{m}$-type

$$R_{\overline{m}}(p) = \begin{cases} p(\bar{\mu}s + (1-\bar{\mu})1) + (1-p)s & \text{for } q < s/r \\ p(\bar{\mu}s + (1-\bar{\mu})1) + (1-p)1 & \text{for } q > s/r \end{cases} \tag{6.15}$$

$R_{\underline{m}}(p) - R_{\overline{m}}(p) > 0$ holds good for $q > s/r$ if

$$p > \frac{1-r}{\bar{\mu}(1-s)+(r-s)(\underline{\mu}-1)} \tag{6.16}$$

$R_{\underline{m}}(p) - R_{\overline{m}}(p) > 0$ holds good for $q < s/r$ if

$$\frac{\underline{\mu}}{1-\bar{\mu}} > \frac{1-s}{r-s} \tag{6.17}$$

Note that (6.16) emerges only for $\bar{\mu}(1-s) > (r-s)(1-\underline{\mu})$ or $(1-s)/(r-s) > (1-\underline{\mu})/\bar{\mu}$ otherwise the relation would be reversed with a negative right-hand side requiring a negative p. The relations (6.16) and (6.17) can be fulfilled both for parameter constellations with

$$\frac{1-s}{r-s} \in \left(\frac{1-\underline{\mu}}{\bar{\mu}}, \frac{\underline{\mu}}{1-\bar{\mu}}\right)$$

Only then can the presence of informed I′-types drive a population share $q_t < s/r$ beyond the threshold beyond which the uninformed U-types have good reason to always trust. $R_{\underline{m}}(p) - R_{\overline{m}}(p) < 0$ applies if, in case of $q > s/r$, (6.16) is reversed. In particular this will always be the case for $(1-s)/(r-s) < (1-\underline{\mu})/\bar{\mu}$. For this parameter constellation growth of the population share q of trustworthy individuals is possible only over the range $[0, s/r)$. Of course, $R_{\underline{m}}(p) - R_{\overline{m}}(p) < 0$ also holds good if in case $q < s/r$ relation (6.17) is reversed to $\underline{\mu}/(1-\bar{\mu}) < (1-s)/(r-s)$.

Again, after defining for convenience

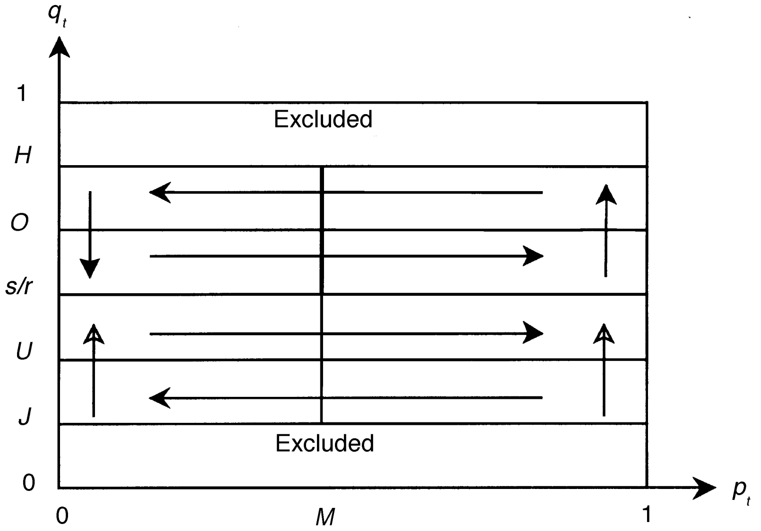


Figure 7 Equation (6.17)

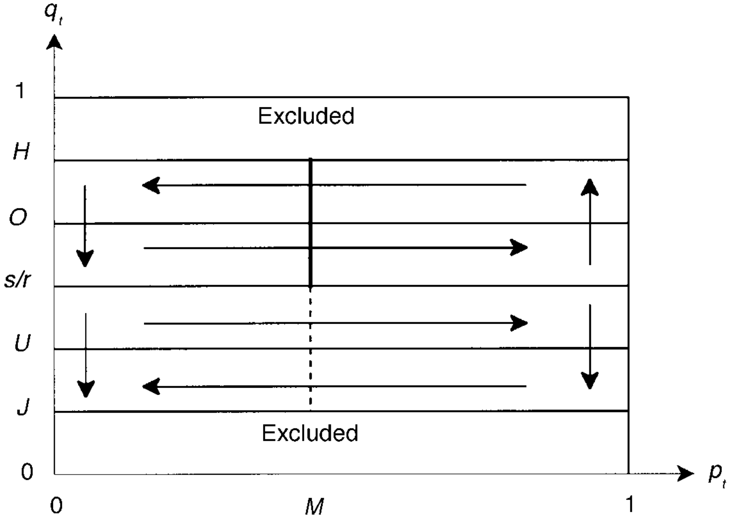


Figure 8 Equation (6.17) reversed

$$M := \frac{1-r}{\bar{\mu}(1-s)+(r-s)(\underline{\mu}-1)} \tag{6.18}$$

our considerations can be summed up graphically by Figures 7 and 8. Note that we restricted attention to parameter constellations satisfying $2C < (s/r)(r-s)(\underline{\mu}+\bar{\mu}-1)$. As a final step let us now turn to the dynamics in $E = \{(p,q) \in [0,1]^2 | q > s/r\}$ if this set contains a unique rest point. We assume that $p > 0$ and $(1-s)/(r-s) > (1-\underline{\mu})/\bar{\mu}$ obtain. Owing to $q > s/r$ we can, as before in the case of differential equations (5.5) and (5.6), restrict attention to two generalized differential equations:

$$\dot{p}_t = k[R_{\mathrm{I}'}(q) - R_{\mathrm{U}}(q)] = k[\{\underline{\mu}(r-s) + (1-\bar{\mu})s - r\}q_t + \bar{\mu}s - 2C] \tag{6.19}$$

$$\dot{q}_t = h[R_{\underline{m}}(p) - R_{\overline{m}}(p)] = 1[\{\bar{\mu}(1-s) + (r-s)(\underline{\mu}-1)\}p_t - 1 + r] \tag{6.20}$$

Making the same basic assumptions about the characteristics of the dynamic process $\dot{p}_t$ again emerges as an (affin) linear function of q_t and $\dot{q}_t$ as an (affin) linear function of p_t. As in the extreme parameter constellation $\bar{\mu} = \underline{\mu} = 1$, the unique rest point

$$(p^*, q^*) = (M, O) = \left(\frac{s\bar{\mu} - 2C}{(1-\underline{\mu})(r-s) + \bar{\mu}s}, \frac{1-r}{\bar{\mu}(1-s) + (r-s)(\underline{\mu}-1)}\right) \tag{6.21}$$

of E has merely the degenerate attraction set $\{(p^*, q^*)\}$. All other starting points $(p,q) \in E$ lead to indefinite cycling around the unique rest point (p^*, q^*).

Comparing the results of Sections 4 and 5, in which perfect type detection was assumed, with those of Section 6 shows that the central qualitative results carry over from the deterministic case characterized by $\bar{\mu} = \underline{\mu} = 1$ to stochastic cases characterized by imperfect type detection. We can thus answer the obvious query whether the assumption of perfect type detection might be so extreme as to render irrelevant all results reached under this premise. There is a clear continuity between the extreme idealization of perfect type detection and the more realistic case of imperfect type detection such that the extreme can be approximated by the less extreme case. On the other hand, basic insights gained by studying the somewhat more transparent extreme case carry over to less idealized situations. However, certain differences remain and it is instructive to look briefly at some of them in a 'comparative statics' kind of way.

It has already been pointed out that in the case of perfect type detection the 'length of the cycle' around the rest point

$$(p^*, q^*) = \left(\frac{1-r}{1-s}, \frac{s-2C}{s}\right)$$

depends only on s: the closer s approaches $\frac{1}{2}$, the shorter the cycle. The first coordinate $(1-r)/(1-s)$ of the rest point will decrease towards 0 if r increases towards 1 and increases towards 1 if r decreases towards s. The second coordinate $[s-2C]/s$ depends on s and C and will increase towards 1 if C decreases towards 0 while approaching s/r if C increases towards $(s/2r)(r-s)$. Moreover, in (5.5) and (5.6) only the parameter s influences the coefficients of p_t and q_t.

In (6.19) and (6.20) the parameters r, s and $\bar{\mu}, \underline{\mu}$ all enter the determination of the coefficients of p_t and q_t. Likewise the rest point depends on r, s and $\bar{\mu}, \underline{\mu}$, as well as on C. The influence of $\underline{\mu}$ on outcomes decreases if $r - s$ decreases while the influence of $\bar{\mu}$ is strongly linked with s. This illustrates how the analysis of the imperfect type detection case can differentiate between the error of trusting somebody who does not deserve it – which happens with probability $(1-p)(1-\bar{\mu})$ – and the error of failing to trust a trustworthy partner – which occurs with probability $p(1-\underline{\mu})$.

7. CONCLUDING REMARKS

The preceding exercise offers, on the one hand, some interesting insights into methodological problems of rational choice modelling (in a wide sense of that term) and, on the other hand, substantial results about trust relationships. To start with the latter, it is intuitively plausible that the prospects of trustworthy individuals as opposed to untrustworthy ones crucially depend on the existence of individuals who can to some extent discriminate between trustworthy and untrustworthy partners. In our model – provided that the conditions for following the signal are fulfilled – the (costly) information on individuals' trustworthiness leads to a discrimination between types. If the signal indicates a trustworthy second mover this leads to a choice of T in first-mover roles even if $q < s/r$. If the signal received indicates an untrustworthy type, N is chosen even if $q > s/r$. This corroborates the somewhat elementary yet extremely important insight that the stability of cooperation in view of the trust predicament depends at least as much on the presence of the faculty and inclination to discriminate in the role of the trustor as on the disposition to act fairly as a trustee.

In our model discrimination expresses itself – in the realm in which following the signal is rational – by signal-dependent choices of N and T. In social reality it may express itself in somewhat different, though related, forms. In particular, players may be able to choose their interaction partners and thus to discriminate against the untrustworthy by refusing to interact with them in the first place. Moreover, the free choice of partners could include both the option of repeating interactions with the same partner and of ending them. If that is the case, 'the shadow of the future' (Axelrod, 1984) or 'the discipline of continuous dealings' (Smith, 1776/1976) might be operative as well. The presence of this shadow would tend to diminish the value of type discrimination by inducing trustworthy behavior by all types. (For simulation studies of the discipline of continuous dealings in the presence of an exit option see, for example, Schüßler, 1990; Vanberg and Congleton, 1992; and on inducing the untrustworthy to behave as if trustworthy, of course, Kreps *et al.*, 1982.)

This being said and explicitly acknowledging that other factors may discriminate between trustworthy and untrustworthy individuals as well we feel that it is of great importance to go beyond the intuitively plausible and to gain a somewhat better understanding of the impact of alternative information conditions on the relative survival prospects of trustworthy, non-opportunistic and untrustworthy, opportunistic individuals. That such a closer look may yield quite surprising insights is borne out by comparing the results of the present with our former applications of the indirect evolutionary approach to trust problems.

Formerly we have studied the influence of 'C, μ' technologies of different reliability on the population composition under the premise that the players could choose strategically whether or not they would acquire specific type information of reliability μ at cost C. Under this condition, depending on the parameter constellation, either only $p = 0$, or $p = 0$ and $p = p^* > s/r$, would be evolutionarily stable population compositions with attraction sets whose union would comprise the whole interval [0, 1]. In the present model all the basic assumptions are the same with the only exception that the information technology now is not chosen but rather some players are endowed or, for that matter, stuck with this costly technology without that being their own choice. Under both assumptions, that of the strategic choice of the information technology and that of its evolution, bimorphic rest points or evolutionarily stable population compositions that contain opportunistic as well as non-opportunistic types may emerge. However, interestingly enough, at least under very natural and simple assumptions about the dynamics of the evolutionary process, the effects on the population composition are qualitatively very different from the case in which the costs of being informed are subject to a strategic choice: in the model in which information evolves along with preferences there need not be a generic attractor corresponding to the stable bimorphism in the model in which information conditions are strategically chosen.

Relating this to our social experience it seems to suggest that the presence of some 'zealots' who are for some reason or other averse to being exploited or are preoccupied with knowing their partners' true character can have a fundamental impact on the character of social interaction (for a somewhat parallel argument see Coleman, 1983). To put it very bluntly, 'nosiness' may not be such a bad thing altogether. We might even represent it by an intrinsic motivational factor expressing a zealous preference for defending oneself against exploitation even at high extrinsic or objective costs C. Such inclinations may lead to building up the population share of trustworthy individuals in situations where nobody would rationally invest in costly information technologies. For informed individuals will seek out the trustworthy in situations in which the population share of the trustworthy is too low for rendering the choice of T rational. Owing to this, the trustworthy will have a slight differential advantage over the untrustworthy unless this effect is overcompensated by mistaken choices of T that are made without a signal or due to a misleading one.

In fact, even in our former model in which players made a strategic decision about acquiring the information technology the same kind of argument may apply. As long as individuals once in a while invest in a 'C, μ' technology even though it is not rational to do so this – without any zealotry – may be sufficient to secure to the trustworthy an advantage over the untrustworthy. In the presence of such mistaken investments in information $p_t > 0$ will be fulfilled throughout. Again, once in a while trustworthy second movers will be singled out and will be trusted in an environment in which there are too few trustworthy to show trust generally. As in the presence of zealots this will bring about a differential advantage for the trustworthy if it is not overcompensated by mistaken choices of T among those who do not command a specific type of information. The latter may seem a big if, though, if one takes into account how mistakes of different kinds may compensate each other.

Avoiding a formal analysis of mistake-driven evolutionary processes here let us only remark that in those cases in which $p_t > 0$ is sufficiently large to provide a differential advantage for the trustworthy over the untrustworthy a self-supporting process of the following kind may be operative: initially there are too few trustworthy around to make seeking them out worthwhile. But after a while that may change and the informed may have an advantage over the uninformed and thus their population share might increase as well.

Of course, as our model shows, after going beyond a certain threshold the decrease of trustworthy individuals – and for that matter also of informed individuals – will eventually be unavoidable. Still, there might be some institutional means or other to stabilize population shares p of trustworthy individuals once they have grown beyond a certain threshold. This seems to be quite interesting from a policy point of view, too. Admittedly it is speculative, though not completely so since within the framework of our indirect evolutionary approach to the trust predicament it can be shown that the proper workings of a court system depend on the presence of sufficiently many non-opportunistic individuals if adjudicators cannot be selected according to their trustworthiness and thus 'judges are no better than the rest of us' (see Brennan *et al.*, 1997).

Letting our discussion of substantial issues rest with this we may finally turn to some methodological problems. Since we have touched on these issues already in our preceding discussion of the substantial trust problem we can be quite brief. We feel that our models suggest that we explore more carefully what may be regarded as varying 'degrees of indirectness' in the evolutionary approach. Evolutionary models might reach from the extreme of a completely indirect evolutionary approach in which even the faculty to make opportunistically rational choices itself is viewed as endogenous to or emergent from an evolutionary process (see Güth and Kliemt, 1998), over the still extreme more standard direct evolutionary approach in which all behavior evolves without opportunistically rational choices playing a role, to models in which some decisions are modelled as opportunistically rational in the standard sense while the information conditions evolve along with some aspects of preferences, to models in which some information conditions as well as the substantial decisions are subject to opportunistically rational choice while certain aspects of the preferences are emergent, to extreme models in which even the utility function for some game can be chosen strategically (though one may have some second thoughts about that claim as, for instance, made in Frank, 1987). Observing that there is almost a continuum of models of varying degrees of indirectness it seems clear that further research should look at the pros and cons of alternative degrees of indirectness. Doing this we may hope to gain not only additional methodological but also some new insights into such eternal problems like that of 'nature and nurture'.

ACKNOWLEDGMENTS

The authors gratefully acknowledge improvements suggested by Karl Schlag and Urs Schweizer.

MATHEMATICAL APPENDIX TO SECTION 5

Denote the interior of E by

$$E^0 := \{(p,q)|0 < p < 1 \text{ and } s/r < q < 1\} \tag{A.1}$$

A function *f*, *f*:[0, 1]→[0, 1], is a smoothing function if:

$$f(0) = 0 \tag{A.2}$$

$$\exists \epsilon, 0 < \epsilon < \frac{1}{2}\left(1 - \frac{s}{r}\right) : x \geq \epsilon \Rightarrow f(x) = 1 \tag{A.3}$$

$$x \in (0,\epsilon) \Rightarrow 1 > f(x) > 0 \tag{A.4}$$

$$f \text{ is continuous and non-decreasing} \tag{A.5}$$

$$\int_0^\epsilon \frac{dx}{f(x)} = \infty \tag{A.6}$$

We can substitute *h, k* in equations (5.5) and (5.6) by smoothing functions f_i and $g_i, i = 1,2,$ to get

$$\dot{p}_t = [s - 2C - sq_t]f_1(p_t)g_1(1 - p_t) \tag{A.7}$$

$$\dot{q}_t = [(1-s)p_t - (1-r)]f_2\left(q_t - \frac{s}{r}\right)g_2(1 - q_t) \tag{A.8}$$

Proposition

For all solutions $(p_t, q_t)_{t\geq 0}$ of equations (A.7) and (A.8) we have:

$$(p_0, q_0) \in E^0 \Rightarrow \forall t : (p_t, q_t) \in E^0$$

To prove the proposition we utilize equations (A.7) and (A.8) to derive

$$\frac{dp_t}{dq_t} = \frac{\dot{p}_t}{\dot{q}_t} = \frac{[s - 2C - sq_t]f_1(p_t)g_1(1-p_t)}{[(1-s)p_t - (1-r)]f_2(q_t - [s/r])g_2(1-q_t)} \tag{A.9}$$

and then separate variables by transforming it into

$$\frac{[(1-s)p_t - (1-r)]}{f_1(p_t)g_1(1-p_t)}dp_t + \frac{sq_t + 2C - s}{f_2(q_t - [s/r])g_2(1-q_t)}dq_t = 0 \tag{A.10}$$

For any $(p,q) \in E^0$

$$U(p,q) := \int_{p^*}^{p} \frac{[(1-s)\bar{p} - (1-r)]}{f_1(\bar{p})g_1(1-\bar{p})}d\bar{p} + \int_{q^*}^{q} \frac{s\bar{q} + 2C - s}{f_2(\bar{q} - [s/r])g_2(1-\bar{q})}d\bar{q} \tag{A.11}$$

is a potential function for (A.10). Owing to the next two observations, (i) and (ii), the function $-U(\cdot,\cdot)$ is single peaked on E^0 with a unique maximum at the point $(p^*, q^*) = ((1-r)/(1-s), (s-2C)/s)$:

(i) The derivative dU/dt of $U(\cdot,\cdot)$ along the line segment

$$\{(p^*, q^*) + t[(p,q) - (p^*, q^*)] | t \in [0,1]\}$$

with $(p,q) \neq (p^*, q^*) is$

$$\begin{aligned}\frac{dU(p(t),q(t))}{dt} &= \frac{\partial U}{\partial \bar{p}}\frac{d\bar{p}}{dt} + \frac{\partial U}{\partial \bar{q}}\frac{d\bar{q}}{dt}\\ &= \frac{(1-s)(\bar{p}-p^*)(p-p^*)}{f_1(\bar{p})g_1(1-\bar{p})} + \frac{s(\bar{q}-q^*)(q-q^*)}{f_2(\bar{q}-[s/r])g_2(1-\bar{q})}, (\bar{p}, \bar{q}) \neq (p^*, q^*)\end{aligned}$$

Since $(\bar{p}-p^*)(p-p^*) \geq 0$ and $(\bar{q}-q^*)(q-q^*) \geq 0$ the function U strictly increases with t. $U(p,q)$ increases along all rays emanating from (p^*, q^*) and the level lines of $U(\cdot,\cdot)$ are bounded in length.

(ii) Because of (A.6) and (A.11) a level line of $U(\cdot,\cdot)$ cannot intersect the boundary of E^0. Moreover, the trajectory of a solution is a level line of $U(\cdot,\cdot)$ since $U(p,q)$ is given as the integral of (A.9) and (A.10) respectively. Thus for any initial point $(p_0, q_0) \in E^0$ the solution will remain in E^0.

As already argued the iso-level curves of $U(p,q)$, i.e. the sets,

$$I(p,q) = \{(\bar{p},\bar{q}) \in E^0 | U(\bar{p},\bar{q}) = U(p,q)\} \tag{A.12}$$

for $(p, q) \in E^0$ are the trajectories $(p_t, q_t)_{t\geq 0}$ of the dynamic system (A.7) and (A.8) for some $(p_0, q_0) \in E^0$. As in the special case of equations (5.5) and (5.6) the system (A.7) and (A.8) with $(p_0, q_0) \in E^0$ is dynamically though not asymptotically stable. Cycling around (p^*, q^*) is in this sense a robust result.

Solution trajectories starting in E^0 remain in E^0 under quite general transformations: Let ϕ and φ be continuous and strictly increasing on $[-1, 1]$ such that $\phi(0) = 0 = \varphi(0)$. Instead of equalities (A.7) and (A.8) we now consider

$$\dot{p}_t = \phi([s - 2C - sq_t])f_1(p_t)g_1(1-p_t) \tag{A.13}$$

$$\dot{q}_t = \varphi([(1-s)p_t - (1-r)])f_2\Big(q_t - \frac{s}{r}\Big)g_2(1-q_t) \tag{A.14}$$

The results stated in (i) and (ii) as derived in our previous discussion of the special case (A.7) and (A.8) carry over to (A.13) and (A.14). Cycling around (p^*, q^*) is thus not brought about by the linearity assumption made in specifying (4.6) and (4.11) respectively, but emerges quite generally when preferences and information co-evolve as described here.

REFERENCES

Arthur, W. B. (1993), 'On Designing Economic Agents that Behave like Human Agents', *Journal of Evolutionary Economics* 3, 1–22.

Axelrod, R. (1984), *The Evolution of Cooperation*, Basic Books, New York.

Brennan, H. G., W. Güth and H. Kliemt (1997), 'Trust in the Shadow of the Courts', Discussion Paper No. 9789, Center for Economic Research, Tilburg.

Coleman, J. S. (1983), 'Free Riders and Zealots', in: W. Sodeur *et al.* (eds), *Ökonomische Erklärungen sozialen Verhaltens*, Verlag der Kooperative, Duisburg, pp. 135–152.

Cressman, Ross, William G. Morrison and Jean-Francois Wen (1998), 'On the Evolutionary Dynamics of Crime', *Canadian Economics Association* 31(5), 1101–1117.

Frank, R. (1987), 'If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?', *American Economic Review* 77(4), 593–604.

Frey, B. S. (1997), *Not Just for the Money. An Economic Theory of Personal Motivation*, Edward Elgar, Cheltenham.

Friedman, Daniel (1991), 'Evolutionary Games in Economics', *Econometrica* 59(3), 637–666.

Fukuyama, F. (1995), *Trust. The Social Virtues and the Creation of Prosperity*, Free Press, New York.

Gale, John, Kenneth G. Binmore and Larry Samuelson (1995), 'Learning to be Imperfect: The Ultimatum Game', *Games and Economic Behavior* 8, 56–90.

Güth, W. and H. Kliemt (1994), 'Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes', *Metroeconomica* 45(2), 155–187.

Güth, W. and H. Kliemt (1995), 'Evolutionarily Stable Co-operative Commitments', *Humboldt University Discussion Paper, Economics Series* 53.

Güth, W. and H. Kliemt (1998), 'The Indirect Evolutionary Approach', *Rationality and Society* 10(3), 377–399.

Hart, H. L. A. (1961), *The Concept of Law*, Clarendon Press, Oxford.

Hofbauer, Josef and Karl H. Schlag (1998), 'Sophisticated Imitation in Cyclic Games', Rheinische Friedrich-Wilhelms-Universität, Bonn. Sonderforschungsbereich 303, Discussion Paper No. B-427.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson (1982), 'Rational Cooperation in the Finitely-Repeated Prisoners' Dilemma', *Journal of Economic Theory* 27, 245–252.

Landa, J. (1994), *Trust, Ethnicity, and Identity*, University of Michigan Press, Ann Arbor, MI.

Posch, Martin (1997), 'Cycling in a Stochastic Learning Algorithm for Normal Form Games', *Journal of Evolutionary Economics* 7, 193–207.

Schüßler, R. (1990), *Kooperation unter Egoisten*, Oldenbourg, München.

Seligman, A. (1997), *The Problem of Trust*, Princeton University Press, Princeton, NJ.

Smith, A. (1776/1976), *An Inquiry into the Nature and Causes of the Wealth of Nations*, Liberty Press, Indianapolis.

Vanberg, V. J. and R. D. Congleton (1992), 'Rationality, Morality and Exit', *American Political Science Review* 86, 418–427.