From Teleology to Evolution[1]

Bridging the gap between rationality and adaptation in social explanation

Siegfried Berninghaus, Werner Güth and Hartmut Kliemt

Abstract

This paper focuses on the uneasy alliance of rational choice and evolutionary explanations in modern economics. While direct evolutionary explanations rule out "purposeful" rational choice by assuming "zero-intelligence" and pure rational choice explanations leave no room for "selective" adaptation the indirect evolutionary approach integrates both perspectives. Subsequently we go stepwise "from teleology to evolution" and thereby study the model spectrum ranging from pure rational choice over indirect to direct evolutionary approaches. We believe that knowledge of this spectrum can help to choose more adequate models of economic behavior that incorporate both teleological and evolutionary elements.

1. Introduction

In neo-classical economics all acts are explained by expectations and evaluations of future effects of action as endorsed by the rational actors themselves. According to the "teleological model" of purposeful action choices are the outcome of the rational pursuit of ends. Since rational actors typically act in the presence of other rational actors they must also form expectations about what these other rational actors do in pursuit of their ends. As far as this is concerned neo-classical economics assumes a strong form of “theory absorption” (Morgenstern, Oskar and Gerhard Schwödiauer 1976, and, in a more philosophical vein Dacey, Raymond 1976) as spelled out in full by modern non-co-operative game theory: The *teleological theory* explaining the actions of rational actors is *commonly known and is in fact applied* by them to choose their own actions. Due to this assumption neo-classical economics becomes a theory of reasoning about action rather than an empirical behavioral science.

In view of the obvious weaknesses of adopting such an "eductive" approach (see Berninghaus, Güth and Kliemt, 2003, for more details) as an empirical theory of choice some

[1] We gratefully acknowledge the very helpful constructive comments, corrections and the encouragement of our referee. Of course, the conventional disclaimer applies.

economists have suggested that the traditional rational choice approach to explaining social behavior be substituted by adaptive models borrowed from (evolutionary) biology and/or from (learning) psychology. More often than not such adaptive modeling leads to the complete elimination of all purposeful choice from the explanation of social behavior. However, as a matter of fact there is purposeful forward-looking choice and thus a teleological element in real world behavior of higher organisms that must not be neglected but be taken into account along with non-teleological springs of action. For a balanced view of human behavior we need both the teleological and the evolutionary (adaptive) perspective.

Subsequently we shall explore the full spectrum of conceivable approaches ranging from those based exclusively on “farsighted teleology” to purely adaptive ones that explain phenomena in terms of “blind evolution” only. For our exploratory purposes we construe a sequence of models each presenting a specific view of the same dyadic "trust"-interaction in a large group of potential partners. In the sequence of models the role of teleology decreases “stepwise”. Starting with the most extreme case in which all choices are explained as “purposeful action” of "rational economic men" we gradually substitute rational choice by "blind evolution" until “purposeful action” and "teleology" are completely eliminated.

More specifically, we consider the following cases of interaction models arranged according to the decreasing degree in which teleology or conventional rational choice assumptions are utilized to explain how trust problems are dealt with:

case a. all crucial elements of the interaction including their own inclination to behave in trustworthy or untrustworthy ways are rationally chosen by the players,

case b. the trustworthiness or untrustworthiness of players evolve whereas everything else is rationally decided,

case c. trustworthiness and the tendency to acquire information about the trustworthiness of others co-evolve,

case d. instead of being fixed strategically all elements of the interaction evolve.

Case a is the obvious starting point since everything that can conceivably be so chosen is in fact assumed to be chosen opportunistically. Since it is somewhat unusual to treat preferences as subject to choice making it seems appealing to consider case b in which a "preference" or "disposition" to behave trustworthy evolves. A possible case b' in which (un-)trustworthiness

– or preferences to that effect – would be rationally chosen while the inclination as well as some faculty to detect such decisions would evolve is not analyzed separately since basic aspects of the evolution of "detection technologies" are covered by the next model of the sequence. In gradually restricting the role of "teleology" or purposeful choice based on rational deliberation studying evolutionary processes in which at least two aspects co-evolve seems more important. The inclination and faculty to detect the presence of commitments to trustworthy courses of action with some reliability is intimately related to the adaptive value of trustworthiness itself. Therefore letting the detection skill co-evolve with trustworthiness and to consider the interdependence between the two adaptive processes as in case c seems a natural next step. In the obvious final step forward-looking strategic choice is completely eliminated from the picture and therefore in case d evolution explains behavioral adaptation without any teleological element.

There are no decisive a priori reasons why any specific model from the spectrum of possible models characterized here should be preferred in principle. Independently of the choice of one specific model basic results are qualitatively very similar. However, even though differences in results cannot tell better from worse models the underlying (behavioral) laws are fundamentally different. In that sense an explanation based on teleology differs dramatically from one based on evolutionary selection. If we are interested in the truth of our theories rather than merely their "predictive" success we must consider to what extent in *fact* either teleological choice or evolution bring about the behavioral adaptations we observe and must express our factual convictions by choosing appropriate models.

For our analysis we make basically three modeling assumptions. Firstly, that the society is large in the sense that conditions akin to anonymity prevail is approximated by the assumption of a pool of infinitely many individuals. Secondly, the presence of both the chance of mutual gain and the risk of default in dyadic relationships is approximated by the assumption that the individuals play basic trust games in which trust can be rewarded or be exploited on each round of play. To reduce some of the complexity and to allow for an analytical treatment it is assumed, thirdly, that individuals cannot strategically choose their partners but are randomly matched to form pairs of players engaged in dyadic interaction.[2]

We hope that after going through the sequence of models a clearer view of the relative merits of rational choice and adaptive modeling of behavior emerges. In section 2 we introduce the

---

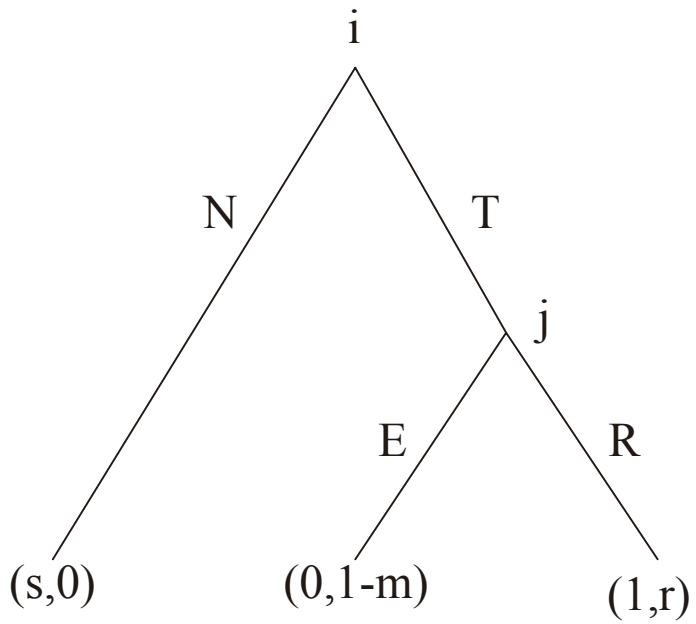
[2] Thus there is no "discipline of continuous dealings" since players are substituted for each other on each round of play.

basic game that we shall subsequently discuss. In section 3 we embed the basic game in a more comprising one and solve the larger game for the extreme case of purely forward looking deliberation or case a. Sections 4 and 5 characterize solutions of the larger game for the intermediate cases b and c respectively. Section 6 discusses the other extreme case d of direct evolution. In section 7 we shall draw some essential methodological conclusions from our "guided tour" reaching "from teleology to evolution". In the appendix, 8, we provide some additional evidence complementary to the basic argument and answer some possible queries.

2. The basic trust game

The example that we use to illustrate the methodological issues at hand starts from what we call the "game of trust" or the "trust game":

Figure 1: The game of trust with payoff parameters $0<r, s<1, m\geq 0$



In the game of Figure 1 player i starts by deciding between *N*(o-trust) and *T*(rust). After *N* the game ends with player i earning s and player j earning 0. After *T* the game continues with j's choice between *E*(xploitation) and *R*(eward). All payoff parameters, except m represent material reward and, in the context of an evolutionary analysis, may be interpreted as measures of "reproductive" success affecting the relative share of different "types" in the population evolving in the course of time.

In an indirect evolutionary approach, subjective and objective payoffs that can differ from each other can both play a role. A subjective payoff function and an objective payoff function apply simultaneously. The former is driving choice the latter selection. The subjective and the objective payoffs are represented by the same numerical payoff function. We assume that except for m, actors are motivated exclusively by the material or objective payoffs involved; i.e. they subjectively evaluate the states of the world according to the emergent states' contributions to their material or objective success. Thus the objective or material payoff function ("reproductive" success) crucial for evolution emerges by setting m=0 after deriving the solution payoffs for all m-types.

If m=0 the individual is exclusively motivated by "extrinsic" or "material" rewards as measured by the corresponding objective success function with the same values. With m≠0 a utility function with motives other than factors directly relevant for evolutionary success emerges. If m is positive, we will often speak of "regret" or the presence of a "conscience". If m>1-r her conscience induces the second moving player j to choose *R* rather than *E*. In that case the conscience is sufficiently strong to become "behaviorally effective". The factor m represents a purely "intrinsic" motive. It is not a measure of objective success but affects that success via potentially influencing behavior.

Behaviorally all values of m with m>1-r are equivalent. Subsequently we will therefore assume that whenever m>1-r applies m is fixed at an arbitrary but specific behaviorally effective $m=\overline{m}>1-r$ that dictates the choice of *R* in the second-mover role in the game of trust. Likewise if m<1-r individuals in the second-mover role will show the same behavior as individuals who are solely motivated by material payoffs. The motives expressed by m are not strong enough to be behaviorally effective. All m<1-r are behaviorally equivalent. This equivalence class is represented by $m=\underline{m}$.[3]

The (game) model in Figure 1 describes the archetype of a one-sided trust situation. Such games of trust are typically embedded in richer social structures. These richer structures lead to more interesting and more complex interactions. We will analyze interactions in which m is assumed to be player j's private information though player i may be in a position to acquire some information about player j's m-type at some cost C.

In the extreme case of purely rational deliberation (the influence of the "shadow of the past" is completely lacking) the m-types will be chosen rationally by the actors themselves. They

---

[3] Even spite as expressed by m<0 will not alter the incentives that are present anyway.

make these choices of their own dispositions by anticipating the future implications of being endowed with a "conscience" leading to "regret" or not (see for a different, less explicit neo-classical discussion of choosing a conscience (Frank 1987)). In the remaining cases of (in)direct evolution m or rather the behavioral dispositions it represents evolve depending on the past differential success of m-types. But let us start with the first extreme case in which the players, in a way, can choose their own type (subjective utility function) operative in the game of trust as embedded in the larger interaction.

3. The pure rational choice approach

3.1. The sequential game

In the one-population interaction envisioned here players are assigned (with equal probability) to the first- or the second-mover role in the final trust game. Since we intend to analyze the same situation in all cases this random move is included already in the discussion of case a where it is spurious. Then the following decision process for the two individuals k=i, j with i≠j unfolds:

Stage 1: Individuals k decide to become either trustworthy by "committing" to $m_k>1-r$ or untrustworthy by "committing" to $m_k<1-r$.

The results of these individual decisions remain private information. But nature provides a signal of the ensuing type distribution. After stage 1 the fraction p of trustworthy $m_k>1-r$ individuals in the population is common knowledge.

Stage 2: Before actually playing the basic trust game of Figure 1 and before knowing whether they shall end up in the first- or second-mover role all individuals k "commit" to become either of "type" *U* – such an *u*ninformed player k_U does not invest in information search about his co-player's m-type – or *I* – such an informed player k_I incurs a cost $C(\geq 0)$ for investing in information search about his co-player's m-type.

Since the solution will not depend on it, it may be left open what players might learn about each other's choice of U or I.

Stage 3: An unbiased chance move decides who becomes first- and who becomes second-mover in the trust game.

Also a stochastic signal revealing the second-mover type with a certain reliability reaches those in the first-mover role who decided on investing in information technology on stage 2.

Without loss of generality let us assume that player i is first and player j is second-mover. If i has decided to become an informed type i_I at stage 2 of the game then i receives a signal M informing him about the trustworthiness of the second-moving player j. M=$\overline{M}$ signals a trustworthy type $\overline{m}$ in the second-mover role. M=$\underline{M}$ signals an untrustworthy type $\underline{m}$. The signal is of reliability $1>\underline{\mu}>1/2$ if originating from an untrustworthy $\underline{m}$-type and of reliability $1>\overline{\mu}>1/2$ if originating from a trustworthy $\overline{m}$-type. Which signal an i_I type receives is decided by a move of chance. If j is an untrustworthy $\underline{m}$-type then with probability $\underline{\mu}>1/2$ the signal $\underline{M}$ will indicate the co-player type correctly to the informed first-moving i_I-type. With probability $1-\underline{\mu}$ an incorrect signal $\overline{M}$ indicating a trustworthy $\overline{m}$-type will be received by i_I. Likewise, with probability $\overline{\mu}$ the signal $\overline{M}$ will correctly indicate the presence of a trustworthy type $\overline{m}$ while with probability $1-\overline{\mu}$ the signal will be $\underline{M}$, indicating an untrustworthy co-player of type $\underline{m}$ even though j is in fact a trustworthy $\overline{m}$-type.

Stage 4: The first-mover i chooses between *N* and *T*. After *N* the game ends. Individual i in the first-mover role receives a payoff of $s-\delta_i C$ and individual j in the role of the second-mover receives a payoff of $0-\delta_j C$;

$$\text{where } \delta_k = \begin{cases} 0 & \textit{in case of } U_k \\ 1 & \textit{in case of } I_k \end{cases} \quad \textit{for any } k \textit{ from the player set}$$

After choosing *T* the game continues.

Stage 5: The second-mover j decides between *E* and *R*. After *E* the game ends with a first-mover payoff of $0-\delta_i C$ and a second-mover payoff of $1-m_j-\delta_j C$. After *R* the game ends with a first-mover payoff of $1-\delta_i C$ and a second-mover payoff of $r-\delta_j C$.

This completes the description of the first model in the sequence. In this initial extreme case "teleology" or purposeful decision-making is extended to the choice of m and information. The prevalence of player types is "explained" or "predicted" solely in terms of (sequentially) rational choices as derivable from the conventional game theoretic logic of solving a sequential game with incomplete information by means of backward induction.

3.2. Solving the game

On stage 5 the second-mover j's decision depends solely on her m-type:

m_j>1-r leads to the choice of *R* and

m_j<1-r leads to the choice of *E*.

On stage 4 we have to distinguish players i_U who lack a specific signal about the second-mover's type and players i_I who have received a signal $\overline{M}$ or $\underline{M}$ conveying specific type information about the second-mover.

The beliefs of a player i_U are determined by the population share p of individuals k who have chosen m_k>1-r on the first stage (in an infinite population[4] the player's private information on her own type is irrelevant). The optimal behavior of a player i_U is to choose *T* if p>s, *N* if p<s.

A player i_I who has received specific type information about her co-player in form of the signal $\overline{M}$ expects a trustworthy $\overline{m}$-type with probability

$$P(\overline{m}/\overline{M}, p)=\frac{p\,\overline{\mu}}{p\,\overline{\mu}+(1-p)(1-\underline{\mu})}.$$

For p≥0 and 1>$\overline{\mu}$, $\underline{\mu}$>1/2 this probability is always well-defined.[5] A player 1_I will choose *T* if P($\overline{m}/\overline{M}$, p)>s and *N* if P($\overline{m}/\overline{M}$, p)<s.

A player i_I upon receiving signal $\underline{M}$ expects (nevertheless) a trustworthy $\overline{m}$-type with probability

---

[4] A finite population-model is analyzed by Güth and Kliemt (1998).

[5] For p=0 case $\underline{\mu}$=1 can be analyzed as the limit of P($\overline{m}/\overline{M}$, p=0) as $\underline{\mu}$ approaches 1.

$$P(\overline{m}/\underline{M}, p)=\frac{p(1-\overline{\mu})}{p(1-\overline{\mu})+(1-p)\underline{\mu}}$$

For $p\geq 0$ and due to $1>\overline{\mu}$, $\underline{\mu}>1/2$ this probability is also always well-defined. Player i_I will choose *T* if $P(\overline{m}/\underline{M}, p)>s$ and *N* if $P(\overline{m}/\underline{M}, p)<s$.

On stage 3 chance assigns players with equal probability to their roles as first- and second-mover, respectively and fixes the signal $M=\underline{M}$, $\overline{M}$ for those players who happen to end up as first-movers and did invest in information. No rational choices of strategic actors are made on stage 3.

On stage 2 individuals choose their informational type U_k or I_k not knowing whether they become first- or second-mover. But they anticipate that specific type information about the co-player type becomes relevant only if they are assigned to the first-mover role. Since they know that the latter happens with probability 1/2 the expected payoff differential $\pi(i_I)-\pi(i_U)$ of an informed, i_I, and an uninformed, i_U, type in the first-mover role must exceed twice the cost C of acquiring specific type information; i.e. the requirement is $\pi(i_I)-\pi(i_U)>2C$.[6]

The difference $\pi(i_I)-\pi(i_U)$ between the payoff expectations of an informed and an uniformed individual in the first-mover role depends on the relation between the probabilities $P(\overline{m}/\underline{M}, p)$, p, $P(\overline{m}/\overline{M}, p)$. In the limiting cases p=1 and p=0 we have $P(\overline{m}/\underline{M}, p)=p=P(\overline{m}/\overline{M}, p)$. All three probabilities are equal and $\pi(i_I)-\pi(i_U)=0$. Obviously nothing can be gained by investing a positive cost C in a signaling technology then.

So let us consider $1>p>0$. Note first that $1>\overline{\mu}$, $\underline{\mu}>1/2$ implies $P(\overline{m}/\overline{M}, p)>p$ – equivalent to $\overline{\mu}>(1-\underline{\mu})$ – and $p>P(\overline{m}/\underline{M}, p)$ – equivalent to $\underline{\mu}>(1-\overline{\mu})$, yielding $P(\overline{m}/\overline{M}, p)>p>P(\overline{m}/\underline{M}, p)$. Moreover, cases $s<P(\overline{m}/\underline{M}, p)$ in which T is chosen even after an untrustworthy second-mover has been signaled and $P(\overline{m}/\overline{M}, p)<s$ in which N is chosen even after a trustworthy second-mover has been signaled can be neglected, since choice will not be affected by receiving signals. In view of the preceding only two possibilities remain in case $1>p>0$

---

[6] If stages 2 and 3 would be exchanged and players would know beforehand which role they would be assigned then they would rationally choose to bear the cost of information if C exceeded $\pi(i_I)-\pi(i_U)$, yet otherwise the analysis would remain the same.

(i) $P(\overline{m}/\overline{M}, p)>p>s>P(\overline{m}/\underline{M}, p)$

(ii) $P(\overline{m}/\overline{M}, p)>s>p>P(\overline{m}/\underline{M}, p)$.

Since specific information about the co-player type will be acquired only if $\pi(i_I)-\pi(i_U)>2C$ and since players will follow the signal we need to consider

in case (i) $[p\,\overline{\mu}+(1-p)(1-\underline{\mu})0] +[p(1-\overline{\mu})+(1-p)\,\underline{\mu}]s - [p1+(1-p)0] >2C$ or

$$p[s+(1-s)\,\overline{\mu}-\underline{\mu}\,s-1]>2C-\underline{\mu}\,s$$

in case (ii) $[p\,\overline{\mu}+(1-p)(1-\underline{\mu})0]+[p(1-\overline{\mu})+(1-p)\,\underline{\mu}]s - s >2C$ or

$$p[s+(1-s)\,\overline{\mu}\,-\underline{\mu}\,s]>2C+s(1-\underline{\mu})$$

In sum, whenever some p with $1>p>0$ fulfills the condition of case (i) or case (ii) investment in detection technology can conceivably pay, while for p=1 and for p=0 it cannot pay to become an informed player at any positive cost $C>0$.

On stage 1 decision-makers know that their decisions on this stage affect payoff only if they end up in the second-mover role on the last stages of the game. Bearing this in mind let us distinguish equilibria in pure and in mixed strategies.

Conceivable *pure equilibria* would be characterized either by p=1 or by p=0. Recall first, that we have seen when analyzing stage 2 that nobody would incur the positive cost $C>0$ of acquiring a technology providing specific type information if p=1 or p=0. Now, if p=1 would characterize equilibrium play then it must be rational for everybody to become a trustworthy $\overline{m}$-type type with the disposition to choose *T* in the final trust-game. Yet, those who would decide on becoming an untrustworthy $\underline{m}$-type would go undetected. Since there are no players with the technology to acquire specific type information all players would choose to trust and untrustworthy types fare better on the last stage. Therefore the stage 1 decision to become a trustworthy $\overline{m}$-type cannot be optimal. The only consistent pure choice behavior on stage 1 is the general choice of becoming an untrustworthy $\underline{m}$-type. Obviously, in a world in which the player type remains private information behavior consistent with the assumption

p=0 is in equilibrium. Therefore the single pure strategy equilibrium is characterized by p=0 and no investment in information.[7]

Turning to mixed strategy equilibria, assume that players on stage 1 choose to become $\overline{m}$-types with positive probability p and $\underline{m}$-types with (1-p), 0<p<1. This fixes the population share of trustworthy $\overline{m}$-types at p. The emerging population share p is known to the players when at stage two of the game they choose to become either U-types – with probability q – or I-types – with probability (1-q). Thereby the share q of informed I-type players in the population is fixed.

Obviously, as long as trembles or occasional mistakes are excluded, there can be mixed equilibria characterized by s>p>0 and q=0. For, if first-movers never trust second-movers – neither intentionally nor by mistake – trust can never be exploited. Since, due to q=0, there are no informed players either who could conceivably find out the trustworthy nothing can discriminate between trustworthiness and untrustworthiness. Both types fare equally well and properly mixed equilibria can emerge. Such mixed strategy equilibria being quite uninteresting let us see whether equilibria with p≥s or q>0 are viable.[8]

For a U-type the payoff expected in the role of first-moving player i is s if p < s and p if p≥s. An I-type expects [p $\overline{\mu}$ + (1-p)(1-$\underline{\mu}$)0]+[p(1 - $\overline{\mu}$)+(1-p)$\underline{\mu}$]s – C. A mixed equilibrium of $\overline{m}$,$\underline{m}$-type choices leading to p with 0 < p < 1 that at the same time allows for a mixture of U, I-choices of informational types with 0<q<1 requires that U-types and I-types fare equally well; i.e. both must expect s if p<s and p if p≥s. The p* for which this is the case is:

---

[7] Factoring in occasional mistakes (see Selten 1975, 1988) of players who choose *T* will yield an advantage for those who chose to become $\underline{m}$-types over those who chose to become $\overline{m}$-types. Allowing for an occasional mistake in the information decision as well, some I-types who incurred the costs of becoming informed will be around. If both mistakes apply there could be a potential advantage. However, the order of magnitude of the joint occurrence of both mistakes will clearly be smaller than that of a single mistake in choosing *T* on the last round of play. Therefore the potential advantage of becoming a trustworthy type and of being trusted by the occasional informed players is smaller than the potential advantage of exploiting the uninformed who are occasionally choosing *T* by mistake.

[8] Since $\neg(p < s \wedge q = 0)$ for positive p and q amounts to p≥s v q>0 we have covered all possible cases after dealing with both $p < s \wedge q = 0$ and p≥s v q>0.

$$p^* = \begin{cases} \dfrac{C + (1-\underline{\mu})s}{\overline{\mu} - (\overline{\mu} + \underline{\mu} - 1)s} & for \quad p < s \\ \dfrac{\underline{\mu}s - C}{(1-\overline{\mu})(1-s) + \underline{\mu}s} & for \quad p \geq s \end{cases}$$

If U-types and I-types both expect s then for the sake of consistency p*<s must be fulfilled. If U-types and I-types both expect p then for the sake of consistency p*≥s must be fulfilled. Obviously, large costs C of information rule out the "mixed" behavior under consideration since neither of the two conditions for p* can be met. If, however, C is sufficiently small then for the equal payoff expectation s of both informational types the condition p* < s is fulfilled due to $\overline{\mu}$ + $\underline{\mu}$ > 1. Similarly, for sufficiently small C, due to $\underline{\mu}$ > 1 - $\underline{\mu}$ for an equal expectation of p the condition p* ≥ s is true.

Assume that C is sufficiently small and that $p^* \in (0,1)$ allows for mixed strategy choices that render choice makers indifferent between becoming an I or U-type. To be in overall equilibrium the mixed U, I-type choices must lead to some $q^* \in (0,1]$, such that players are indifferent between choosing $\overline{m}$- or $\underline{m}$-dispositions. Again the two cases p<s and p≥s must be distinguished.

If p<s all U-types choose N and thereby $\overline{m}$-types and $\underline{m}$-types are treated equally and receive s indiscriminately in the second-mover role. For I-types, that appear with probability q*>0, indifference requires that

$$q^*[\underline{\mu}0 + (1-\underline{\mu})(1-\underline{m})] = q^*[\overline{\mu}r + (1-\overline{\mu})0].$$[9]

As a necessary condition for a mixed equilibrium we can derive from this

$$\underline{m}^* = 1 - \frac{\overline{\mu}}{1-\underline{\mu}}r.$$

Additionally it is required that $\underline{m}^* < 1 - r$ or $\underline{m}^* = 1 - \frac{\overline{\mu}}{1-\underline{\mu}}r$<1-r. The latter, however, is always fulfilled since by assumption $\overline{\mu} > 1/2 \wedge \underline{\mu} > 1/2$ and thus $\overline{\mu} + \underline{\mu} > 1$ or $\frac{\overline{\mu}}{1-\underline{\mu}}$>1. In sum, within the range 0<p<s a mixed equilibrium characterized by $p^* = \frac{C + (1-\underline{\mu})s}{\overline{\mu} - (\overline{\mu} + \underline{\mu} - 1)s}$, $\underline{m}^* = 1 - \frac{\overline{\mu}}{1-\underline{\mu}}r$ can exist for arbitrary $q^* \in (0,1]$.

---

[9] Since p<s uninformed players will never trust and the other equilibrium with (p*, q*=0) would emerge if q*=0 would be allowed for.

If p≥s then the population share of U-types who trust indiscriminately is (1-q*) and untrustworthy $\underline{m}$-types expect from interacting with these individuals in the second-mover role (1-q*)(1-$\underline{m}$) while trustworthy $\overline{m}$-types expect (1-q*)r. Including the expectations of the trustworthy and the untrustworthy stemming from the behavior of the uninformed U-types along with their expectations arising from the discriminating behavior of the informed I-types (as in case p<s) we get

$$q^*[\underline{\mu}0+(1-\underline{\mu})(1-\underline{m})]+(1-q^*)(1-\underline{m})=q^*[\overline{\mu}r+(1-\overline{\mu})0]+(1-q^*)r.$$

Clearly, q*=1 directly implies again $\underline{m}^*=1-\frac{\overline{\mu}}{1-\underline{\mu}}r$<1-r. So let us finally ask whether there can be $q^*=\frac{r-1+\underline{m}}{r(1-\overline{\mu})-\underline{\mu}(1-\underline{m})}\in(0,1)$ for some $\underline{m}$<1-r. Obviously $\underline{m}$<1-r implies r-1+$\underline{m}$<0 and thus, since q*>0 is required also r(1-$\overline{\mu}$)-$\underline{\mu}$(1-$\underline{m}$)<0. From this we get $\underline{m}<1-\frac{1-\overline{\mu}}{\underline{\mu}}r$. Since 1-$\overline{\mu}$<$\underline{\mu}$ or 1<$\overline{\mu}$+$\underline{\mu}$ we have $\underline{m}<1-r<1-\frac{1-\overline{\mu}}{\underline{\mu}}r$. Checking whether for such an $\underline{m}$ we have q*<1 we must bear in mind that for $\underline{m}$<r-1 the denominator (in the equation for q*) is negative which implies that q*<1 amounts to r-1+$\underline{m}$>r(1-$\overline{\mu}$)-$\underline{\mu}$(1-$\underline{m}$). From this we infer $\underline{m}>1-\frac{\overline{\mu}}{1-\underline{\mu}}r$. As can be immediately seen $1-\frac{\overline{\mu}}{1-\underline{\mu}}r<1-\frac{1-\overline{\mu}}{\underline{\mu}}r$. Therefore there is a generic interval such that $\underline{m}\in(1-\frac{\overline{\mu}}{1-\underline{\mu}}r,1-\frac{1-\overline{\mu}}{\underline{\mu}}r)$ and 1>p≥s along with 0<q*<1.

In sum, there can be mixed strategy equilibria where

- $\overline{m}$-dispositions to behave trustworthy and $\underline{m}$-dispositions to behave untrustworthy are both rationally chosen with positive probabilities of p, (1-p), respectively, and
- U- and I-type behavior is rationally chosen with positive probabilities (1-q), q, respectively,

provided that the costs C of becoming an I-type are sufficiently small and $\underline{m}$ fulfills certain other requirements as imposed by the model itself.

4. Choice dimensions as evolving

In the strict rational choice analysis the commitment decision to behave trustworthy at the last stage of the game was itself made strategically in view of the payoffs that it might bring about

to be so committed. Such explicit commitment decisions may be plausible modeling assumptions in certain circumstances. However, the disposition to behave trustworthy or not is often a general trait of character that develops through time (conceivably being even innate). In many instances it is quite implausible to assume that developing a general disposition to behave trustworthy originates in strategic decisions to that effect. Therefore in our first step towards substituting rational strategic decisions by adaptive processes it is not anymore assumed that the individuals choose their own "commitment-type". In the model emerging after the first step of the modification process the population share p_t, $0 \leq p_t \leq 1$ of trustworthy $\overline{m}$-type individuals develops through time (t) in an evolutionary process that is not driven by strategic type-choices but rather by type-selection. Trustworthiness or the parameter $\overline{m}$=m>1-r become "an endowment" of the individual that is fixed by "nature" rather than being strategically chosen by the actors themselves.

It is assumed that except for the parameter "m" subjective utility functions coincide with the material or objective payoff functions that measure success and that the dynamics of p_t are monotonic in objective or material success: The population share p_t of trustworthy individuals increases if, depending on their (subjectively motivated) individual decisions, the trustworthy are more successful than the untrustworthy. Likewise the population share p_t of trustworthy individuals decreases if the untrustworthy are more successful.

"Subjective preferences" and expectations do not directly affect objective success. But the purely subjective factor m alters behavior in the basic game. The population composition, p_t, represents the prevalence of the subjective factor in the population in general and fixes commonly known priors. The general knowledge of p_t along with specific information about the type, m, of a co-player in the second-mover role will influence rational expectations and rational choices that in turn influence objective success and consequently p_t.

We use a function $R(p_t, m)$ to measure objective success as depending on the determinants of behavior:

$$p_t \quad \begin{cases} increases\ with \quad t \quad if \quad R(p_t, \overline{m}) > R(p_t, \underline{m}) \\ decreases\ with \quad t \quad if \quad R(p_t, \overline{m}) < R(p_t, \underline{m}) \end{cases}$$

Success is determined by solving the ("remaining") strategic game G. The "rules of that game" are dependent on the m-types m_i, m_j of the two randomly matched individuals i and j and on p_t such that a class of games each of the form $G(p_t, m_i, m_j)$ emerges. Let $s^*(p_t, m_i, m_j)$ be the solution of $G(p_t, m_i, m_j)$. Assuming that $s^*(p_t, m_i, m_j)$ is unique we can treat the

objective success measure R and the subjective evaluation or utility u as functions of (p_t, m_i, m_j). Then the vector u(s*(p_t, m_i, m_j)) forms the solution payoff vector in subjective (motivationally relevant) terms and R(s*(p_t, m_i, m_j))=R(p_t,m_i,m_j) is the solution payoff vector in objective (directly evolutionarily relevant) terms. The values of R are determined in two steps:

In **step 1** of the indirect evolutionary analysis the solutions of a class of Bayesian games must be derived. The Bayesian games emerge after substituting the strategic stage 1 type-choice of the pure rational choice game of the previous section by a fictitious random move that determines player-type according to the population composition p_t before each round of play t. More specifically, at time t any player will with probability p_t be of the trustworthy type $\overline{m}$ and with probability (1-p_t) of the untrustworthy $\underline{m}$-type. As far as subjective expectations about the degree of trustworthiness in the population at large are concerned we assume that at each point in time t the parameter p_t is common knowledge among the players. Since the population is assumed to be infinite the emerging a priori beliefs of players about the co-player type do not depend on their private knowledge of their own type. Under these conditions for any t the solution s*(p_t, m_i, m_j) of each of the emerging games G is derived exactly as in the preceding section.

In **step 2** of the indirect evolutionary analysis the results of step 1 are used to determine whether the share of trustworthy individuals increases or decreases. This in turn determines the evolution of p_t through time. Since in step 1 the rational choice solution has been determined for each of the games from the full class of games emerging for each $0 \leq p \leq 1$ we know the objective success corresponding to the solution payoffs for $\overline{m}$- and $\underline{m}$-type players respectively. Therefore we can say which of the types at each state of evolution outperforms the other one. The first step describes the motivational factors or the subjective side of the matter in a more or less traditional framework of rational forward-looking decision-making while the second step is of the completely different kind of "blind" selective adaptation. Combining the two steps in one model we have entered the middle ground between “teleology and adaptation”.

A typical question to be asked here in the spirit of “comparative statics” concerns the so-called evolutionary stability of population compositions (for additional details of this see (Güth and Kliemt 2000)). As in case of the "mixed" equilibria of the preceding (pure) rational choice model the results of adopting such a method in the circumstances envisioned here

depend on the costs of gaining specific information and on the reliability of the available information technology. Consider the following figure:

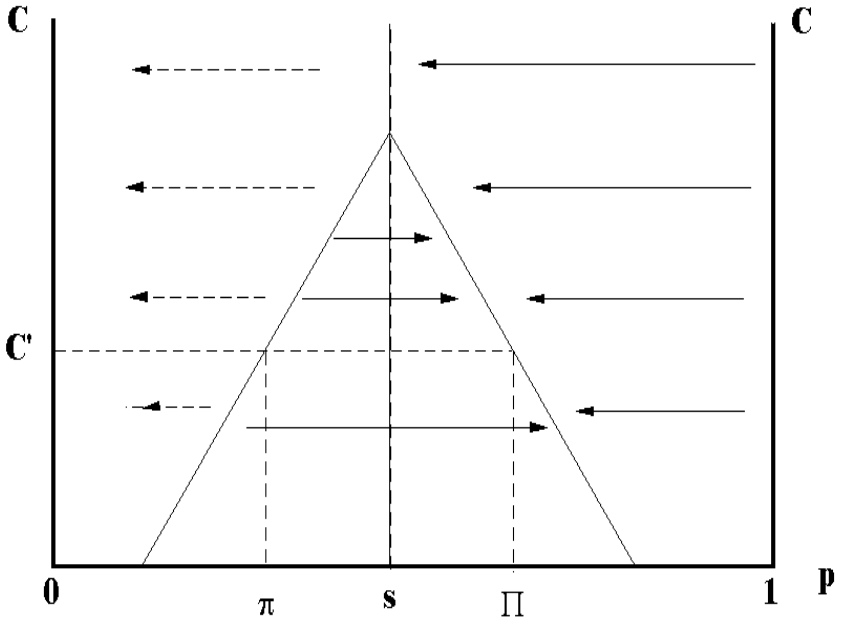


Figure 2: The adaptation of p_t over time

In Figure 2 the horizontal axis represents all possible shares p of trustworthy $\overline{m}$-types in the population. For any initial population composition parameter p_0 the figure illustrates the development of p_t through time if a detection technology of reliability $\underline{\mu}$, $\overline{\mu}$ to find out $\overline{m}$ and $\underline{m}$-types, respectively, is available at cost C. For each C solid or dashed arrows show the direction in which p develops if the initial parameter p_0 lies somewhere on the arrow's starting line. Dashed arrows indicate an evolutionary advantage that comes about only if individuals once in a while make mistakes in their choices by deviating from what rationality dictates. Solid arrows show the direction of evolution if under fully rational behavior (in step 2 of the analysis) certain types have the advantage over other types. The shape of the triangle is fixed by the reliability of the available technology and the objective payoff structure.

The vertical line starting at s indicates the threshold value for the decision to trust or not to trust in the basic game if only the population composition p is known, corresponding to a situation outside the triangle of Figure 2. In that case rational first-movers will show trust if $p \geq s$ and show no trust if $s > p$. Within the triangle around the s-line we see how the presence of a technology for gathering specific information about the co-player type affects the population composition. Here beyond the parameter p that characterizes the population at large further

information about the specific co-player who is assigned to act in the second-mover role is acquired at cost C. For each population composition p_0 that is for some cost level C located in the triangle the population composition parameter p will grow until the right border line of the triangle is hit at cost level C. In this realm individuals have an incentive to reach a positive decision to become informed (and the presence of informed individuals makes it potentially advantageous to be trustworthy).

More generally, for each sufficiently small value of C we get an interval (around s) of population compositions p for which it pays to invest in the information technology. Investing in specific type information pays if the probability that this specific information leads to a beneficial alteration of behavior is high enough. This probability depends on the reliability of the information technology and on the population composition.

To start with the latter, assume for instance that $p<s$ is from the relevant interval for some sufficiently low C (i.e. it lies within the triangle). Then the uninformed individual who did not invest in information technology would play N on the final round. But it is worthwhile (or cheap enough) to invest in the technology that yields specific information telling with some (sufficient) reliability whether a second-mover deserves to be trusted. If after investment a specific signal of trustworthiness is received a move other than the one dictated by the knowledge of the general population composition will be made. Since making this move selectively leads to increases in expected gains beyond investment costs, informed individuals even though they incurred C have the edge over the uniformed.

Intuitively it should be clear also that for population compositions close to p=0 it will in all likelihood not pay to invest in a costly information technology to find out those second-movers who – regardless of the population composition suggesting N – would deserve to be trusted. Close to p=0 there are simply too few trustworthy around to make it worthwhile to seek them. Only if perfectly reliable information were available at no cost it would be worthwhile to acquire specific information for all population compositions $p>0$. Vice versa, for compositions close to p=1 it will not pay to invest $C>0$ to find out the individuals who should not be trusted even though the commonly known population composition parameter would suggest the choice of T. If too few untrustworthy are around, it does not pay to bear the cost of finding them.

Summing up this line of argument it is obvious that for suitable values of $\underline{\mu}$, $\overline{\mu}$, C the interval [0, 1] of possible population compositions can be divided into three sub-realms $(0, \pi)$, $(\pi, \prod)$, $(\prod, 1)$, with $0<\pi<s<\prod<1$. Consider the following initial values:

$p_0 \in (0,\pi) \Rightarrow [t \to \infty \Rightarrow p_t \to 0]$ It does not pay to invest in information technology since there are not sufficiently many trustworthy individuals to make it worthwhile to find them. In view of p<s all players intend to play N. All types would fare equally well then would not occasional mistakes or unintended choices of T offer a differential advantage to the untrustworthy $\underline{m}$ -types.

$p_0 \in (\pi,\Pi) \Rightarrow [t \to \infty \Rightarrow p_t \to \Pi]$ It does pay to invest in information technology in (π, $\prod$). There are sufficiently many trustworthy and there are sufficiently many untrustworthy to make it worthwhile to find them. Since trustworthy individuals are trusted with higher probability than the untrustworthy the trustworthy $\overline{m}$ -types earn a higher material payoff than the $\underline{m}$ -types if p $\in$ (π, $\prod$).

$p_0 \in (\Pi,1) \Rightarrow [t \to \infty \Rightarrow p_t \to \Pi]$ It does not pay to invest in information technology and thus to be able to find out the few untrustworthy individuals. Since, except for occasional mistakes, all types indiscriminately choose T in the first-mover role untrustworthy $\underline{m}$ -types are more successful.

Depending on the initial population composition, we will observe that the population dynamics either will eventually decrease to a p=0 population share or converge – from below or from above – to p=$\prod$. These are the only outcomes that will emerge under plausible monotonic dynamics and at the same time the only evolutionarily stable population compositions.

Qualitatively these results are quite similar to those of the "pure" rational choice model. The analogy between the strict equilibrium point of the fully teleological and the rest point p=0 (with its generic attraction set) of the partly evolutionary model is obvious. The same holds good for completely mixed equilibrium solutions in the fully strategic context and for bi-morphisms in the evolutionary context. But regardless of such similarities of their possible outcomes there are clear differences between the models. For instance, in the case at hand it is highly implausible that humans as a *matter of fact* might be in a position to strategically choose their own dispositions of trustworthiness or untrustworthiness. It is much more plausible that, contingent on objective success, such personal characteristics evolve in some adaptive process or other.

Whether or not that is indeed the case is a factual issue that must be decided empirically on grounds other than qualitative differences in predicted results. It should be noted carefully that the qualitative analogy of predicted results does not render that decision unimportant. Quite to the contrary the nature of the explanation depends crucially on the causal laws assumed to

apply. Adherents of rational choice modeling who allow only purposeful strategic choice as basis of their explanations cannot decide that issue by reasons a priori. They must show by independent evidence that choice is in fact purposefully rational. Moreover, contrary to economic folklore it is not sufficient that results are "as if" brought about by purposefully rational choice. That theories are merely instruments for the prediction of results expresses merely a grossly mistaken methodological view. Counterfactual or so-called "potential explanations" are not based on true behavioral laws. Yet in explaining social reality the ultimate aim must be true rather than merely potential explanations. Adequate explanations must be based on those laws that were *in fact* or causally operative in bringing about the results we observe. In so far teleological and adaptive elements as a matter of fact do both play a role in social reality none should be ruled out by modeling assumptions.

The great advantage of the indirect evolutionary approach is that it allows for a "mixture" of nomological assumptions that can pay due respect to teleology and evolution in an integrated model. The preceding intuitive illustration demonstrates how a single adaptive dimension can be included in a standard rational choice model. From this we know in principle how to proceed in cases in which several dimensions are treated in terms of rational choice while a single one is subject to an evolutionary process.[10] How, in principle, more than one (possibly several) adaptive dimension(s) can be included along with rational choice dimensions in the same model is illustrated next (see for a detailed analysis of the following (Güth, Kliemt, Peleg 1999)).

5. Trustworthiness and informational status as co-evolving

Assume that individuals do no longer decide on becoming either informed or uninformed. Rather the informed and the uninformed are selected according to their relative success. The population share q of informed I-types (as opposed to the share (1-q) of uninformed U-types) co-evolves together with the population share p of the trustworthy $\overline{m}$-types (as opposed to the

[10] For instance, we could discuss along the same lines the case where only the presence of the information technology evolves. But as mentioned before this case of evolution along a single dimension is from a substantial point of view not as interesting as the one analyzed here and would not contribute additional insights.

untrustworthy $\underline{m}$-types). That is, we get a population share for each of four possible types: I-$\overline{m}$-type, U-$\overline{m}$-type, I-$\underline{m}$-type, U-$\underline{m}$-type.

Let us refer to the U and I type of individuals as their "informational" and to the $\overline{m}$ and $\underline{m}$ type as their "moral" type. As is obvious from the preceding analyses, play in the final trust game is determined solely by the moral type of second-moving player j and the informational type of first-moving player i. For instance, a trustworthy uninformed U-$\overline{m}$-type and a trustworthy informed I-$\overline{m}$-type both behave exactly the same way in the second-mover role. Therefore behavior in that role cannot imply differential payoffs that discriminate between U-$\overline{m}$-type and I-$\overline{m}$-type. Likewise in the first-mover role moral type does not matter. Two individuals of the same informational type regardless of their moral type behave equally in the first-mover role and therefore must fare equally in that role. Thus, in a sense the two dimensions of the problem can be separated and in spite of the presence of four different types the co-evolutionary process will be two-dimensional only.

Still, what happens along one dimension influences what happens along the other. For instance, whether becoming an informed rather than uninformed individual is worth its costs crucially depends on the proportion p of individuals who are trustworthy. Likewise, whether or not the trustworthy fare better than the untrustworthy depends on the proportion of informed as opposed to uninformed individuals. Only if first-moving individuals command the faculty to single out trustworthy second-movers with sufficient reliability and thereby discriminate against the untrustworthy can it be a differential advantage to become trustworthy (and thereby to incur the opportunity cost of foregoing the chance of exploiting first-mover trust). In sum, the "evolutionary climate" in which the informed flourish is provided by the presence of trustworthy and untrustworthy types in the "right proportion" (not too few and not too many trustworthy individuals must be there) while the trustworthy will flourish the better the more informed individuals are around. So what we should expect is that differential "reproductive" success of the trustworthy in comparison to the untrustworthy will depend on the share of informed individuals, q, while the share of the trustworthy, p, determines whether the informed will fare better or not than the uninformed.

In formal terms we move from considering the dynamics of p in the space [0, 1] to considering the dynamics of p,q-constellations in p,q-space or the unit square formed by the Cartesian product [0, 1]x[0, 1]. In the co-evolutionary process types are selected by their relative success as measured in objective terms. To determine relative success we must again solve the game. The Bayesian games to be solved in step 1 of the indirect evolutionary approach for each of the p,q-constellations are simpler. But step 2 becomes somewhat more

complicated. Fortunately it suffices for our present purposes to consider a rather special case and to convey a more or less intuitive impression of how the (p_t q_t)-dynamics unfolds through time for initial constellations (p_0 q_0). More specifically, let us turn to this task assuming that a perfect information technology providing signals $\overline{M}$ and $\underline{M}$ that are perfectly type-discriminating is available at a positive cost; i.e. $1=\overline{\mu}=\underline{\mu}$, C>0.

If in one p,q constellation the I-$\overline{m}$-type, U-$\overline{m}$-type, I-$\underline{m}$-type, U-$\underline{m}$-type, respectively, is relatively more successful than its competitors it will spread and if it is less successful its share will decrease. We assume that such simple monotonic dynamics prevail and can be represented by linear differential equations[11]. Since informational type only matters in the first-mover role and moral type only in the second-mover role and since the two informational and the moral types are therefore independent, this will translate to increases and decreases respectively along the two dimensions of p and q. What is going on along these dimensions at each point in time, t, is captured by two differential equations:

$$\dot{q}_t = k\,[R_I(p_t)-R_U(p_t)],$$

$$\dot{p}_t = h\,[R_{\underline{m}}(q_t)-R_{\overline{m}}(q_t)],$$

where h, k>0 are positive constants. The first differential equation describes the relative success of the informed as compared with the uninformed the second the relative success of the trustworthy as compared with the non-trustworthy. The equations also clearly illustrate how at each point in time t the change $\dot{p}_t$ of the share p_t depends on q_t and how the change $\dot{q}_t$ depends on the share p_t.

The success function for the informed individuals in the first-mover role is given by $R_I(p) = p + (1-p)s - 2C$ since they detect all the trustworthy players with whom they are matched with probability p corresponding to the population share of the trustworthy. Trusting precisely the trustworthy they receive p+(1-p)s at cost C (which must be doubled since they are assigned to the second-mover role only with probability 1/2).

The success function of uninformed individuals depends on whether p>s or p<s applies. Since the payoff from being exploited is 0 it is obviously given by

---

[11] Linearity allows a closed form description of the process (see Güth, Kliemt and Peleg, 1999). Merely requiring monotonicity would yield the same rest point but may render it asymptotically stable (in the case of linearity the population composition cycles around the mixed rest point).

$$R_U(p) = \begin{cases} s & for\, p < s \\ p & for\, p \geq s \end{cases}$$

The difference between the payoffs of the informed and uninformed is

$$R_I(p_t)\text{-}R_U(p_t) = \begin{cases} p(1-s) - 2C & for \quad p_t < s \\ (1-p)s - 2C & for \quad p_t \geq s \end{cases}$$

Likewise we can derive the payoffs of the two moral types as depending on the prevalence of informed types in the population. All informed individuals are in command of perfect type detection faculties and shall thus choose to trust if and only if the second mover is in fact trustworthy. Since in the second-mover role they will earn 0 if p<s this yields qr as an expectation of the trustworthy in that case. Otherwise, if p>s, everybody will trust the trustworthy who therefore receive r in each and every instance. So depending on q the trustworthy will receive

$$R_{\bar{m}}(q_t) = \begin{cases} qr + (1-q)0 & \text{for } p < s \\ r & \text{for } p \geq s \end{cases}$$

The untrustworthy receive 0 in the second-mover role if p<s. In that case nobody trusts them since the uninformed choose not to trust anyway while the informed will single them out as not worthy of their trust. Analogously in case p>s the untrustworthy shall flourish better due to the trust shown to the undeserving by the uninformed. These considerations amount to

$$R_{\underline{m}}(q_t) = \begin{cases} 0 & \text{for } \quad p < s \\ q0 + (1-q)\,1 & \text{for } \quad p \geq s \end{cases}$$

Finally

$$R_{\bar{m}}(q_t) - R_{\underline{m}}(q_t) = \begin{cases} rq_t - 0 & \text{for } p_t < s \\ r - (1 - q_t) & \text{for } p_t \geq s \end{cases}$$

The preceding remarks should suffice to get an intuitive grasp of what is going on in the co-evolutionary process. Looking at the world through the window of a model of the co-evolutionary process we can study how the populations of alternative types along each of the dimensions are providing evolutionary niches, or not, in which the corresponding other type can flourish or not.

As observed the growth and decline of both types interact with each other in an orderly manner described here by rather specific differential equations. In such analyses naturally the question of rest-points of the evolutionary dynamics emerges. The process comes to a stop or a rest point (p*, q*) if both $\dot{p}_t$=0 and $\dot{q}_t$=0; i.e.

$$(\dot{p}_t, \dot{q}_t) = (k\,[R_I(p^*) - R_U(q^*)],\ h\,[R_{\bar{m}}(q^*) - R_{\underline{m}}(q^*)]) = (0, 0).$$

It can be shown relatively easily that there is a single rest-point with

$$(p^*, q^*) = \left(\frac{s - 2C}{s},\ 1\text{-}r\right) \text{ in the range } E := \{(p, q) \in (0, 1)^2\},$$

while for p<s only the line segment [q=0, p<s) contains rest points.

But the unique mixed rest-point (p*, q*) is not locally asymptotically stable for the specific differential equations at hand. There are no small neighborhoods of the point such that for all starting points from the neighborhood the dynamic process will converge towards the rest point. Therefore the rest point is not a likely candidate for a stable state that might be expected to prevail for any extended period of time. It is rather to be predicted that the dynamics will cycle in a manner illustrated by the graph in Figure 3. Of course, one must also exclude that the process leaves the unit square (see Güth, Kliemt and Peleg, 1999, who show how this can be guaranteed):

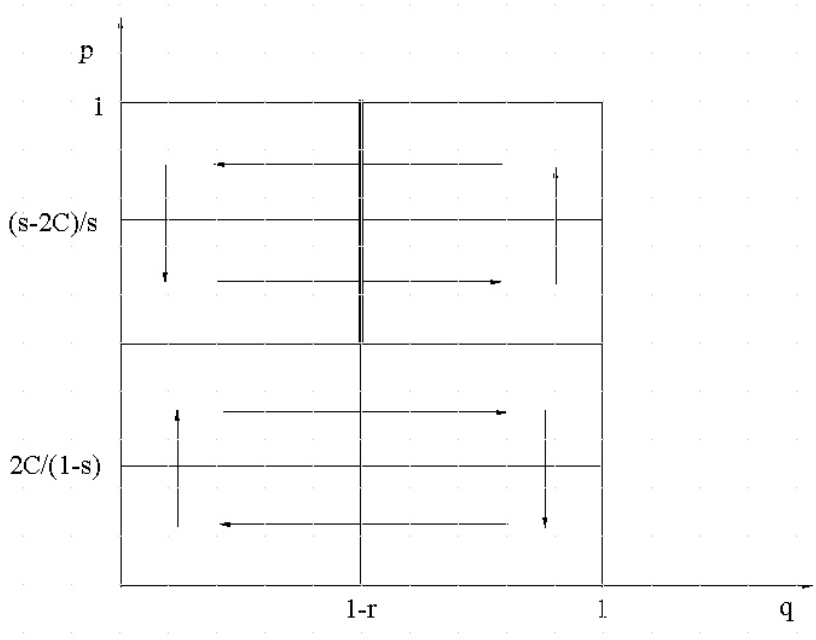


Figure 3: Graph of cycling dynamics

Again analogies between the results of the co-evolutionary process as derived in this section and the results of the two previous models are quite obvious. We may note first that there is a correspondence between the role of the line segment [q=0, p<s) in the present and the attractor p*=0 of the population composition parameter and the pure equilibrium with p=0, respectively. Secondly, the mixed rest point, the stable bi-morphism and the completely

mixed strategy equilibrium are qualitatively similar results derived from each of the three different models. Cycling around the rest point that might emerge in the third model only is, however, a new result of a qualitatively different nature.[12] We do not view such cycling as necessarily counterintuitive. Societies might fluctuate in the sense that phases of high trustworthiness and decreasing mistrust (represented by less I-types) are followed by phases of low trustworthiness and increasing mistrust.

6. Direct Evolution

Up to now it has been assumed that the choice between N and T on stage 4 of the game is a strategically rational one. Going all the way to a completely non-strategic approach this choice is now modeled, too, as resulting from fixed behavioral inclinations that emerge from an evolutionary process.

Whether the trusting action T is chosen has been determined in the preceding models as the outcome of rational strategic choices in the light of the available information. If the action T is to be understood now as the outcome of a fixed behavioral program then the share of individuals in the population who in the first-mover role would show trust, T, can be subject to an evolutionary process, too.

That the behavioral program is fixed does not imply that behavior needs to be unconditional. In higher organisms behavioral programs are triggered by "information" on states of the world as retrieved by the organisms. In the case at hand the behavior shown should depend on whether the individual k is a U_k or an I_k-type. For an I_k-type it is obvious to assume that the signal $\overline{M}$ triggers the choice of T and the signal $\underline{M}$ the choice of N. We thus have to distinguish the population share q of individuals k with I_k and, for the complementary population share 1-q, the sub-share u of those who, as uninformed players, would rely on T.

The dynamics of the shares p_t, q_t and u_t over time are, as before, determined by the difference in reproductive success. Again the dynamics of q_t and u_t are only shaped by differences of reproductive success in the role of the first-mover. Therefore three dimensional dynamics are sufficient. Since U_k-individuals k receive s if they rely on N and p_t if they use T at time t, the

(again linear) dynamics of q_t are determined as

---

[12] Note, however, that overshooting could also cause some cycling around the bimorphism in Figure 2.

$$\dot{q}_t = k[\{p_t\,\overline{\mu} + (p_t(1-\overline{\mu}) + (1-p_t)\,\underline{\mu})\,s - 2C\} - \{u_t p_t + (1-u_t)s\}]$$

by the difference in the reproductive success of I_k- individuals and the weighted reproductive success of U_k-individuals at time t.

For u_t the obvious dynamics are

$\dot{u}_t = a[p_t - s]$ with a>0.

We adjust our preceding discussion of the co-evolutionary process as follows

$$\dot{p}_t = h\,[R_{\overline{m}}(q_t,u_t) - R_{\underline{m}}(q_t,u_t)] \qquad \text{with h>0.}$$

Here $R_{\overline{m}}(q_t,u_t)=[q_t\,\overline{\mu} + (1-q_t)u_t]$r and $R_{\underline{m}}(q_t,u_t)=q_t(1-\underline{\mu})+(1-q_t)u_t$.

Even in such a complex evolutionary setting results can be derived analytically though it is more demanding to do so in a general way and to characterize the evolutionary process completely. But central points can be summarized quite briefly. As shown in the appendix the rest points (p*, q*, u*) of the adaptive process can be characterized in the following way:

> If $(\overline{\mu} + \underline{\mu} - 1)(1-s)s < 2C$ the rest points (p*, q*, u*) are related to "monomorphic" untrusting and uninformed behavior in a world in which trustworthiness is rather rare. They are of the form p*<s, q*=u*=0 and, as simulations show, convergence to such rest points is either monotonic or cyclic.
>
> If $0 < 2C < (\overline{\mu} + \underline{\mu} - 1)(1-s)s$ then in addition to the rest points (p*, q*, u*) with p*<s, q*=u*=0 we have rest points characterized by poly-morphic behavior with u*=1, s<p*<1, 0<q*<1. Simulations show that the additional rest points will not be reached since the process runs in stable cycles around them.

Again results of "blind evolution" are quite analogous with the results that emerge in the presence of some teleological element. We shall come back to this basic point immediately in our final discussion.

7. Discussion

By way of examples we surveyed the full field of approaches reaching from

- pure rational choice assuming as in traditional (neo-classical) economics that there is a shadow of the future but none of the past to

- direct evolution assuming as in traditional (evolutionary) biology that there is only a shadow of the past but none of the future.

The view that rational choice behavior and evolutionary stability amount to the same since evolutionary stability implies the best reply-property is rather naïve and unjustified except for special situations (see, for instance, Güth and Peleg, 2001). Nevertheless, as our survey of models shows, even the extremes of our spectrum, the cases of pure teleology and completely blind evolution, respectively, seem to imply qualitatively analogous conclusions. This provides an explanation for the otherwise surprising fact that the "predictions" of the rational choice approach are often well in line with observations. At the same time, the insight that behavioral assumptions reaching from omniscient teleology to blind evolution can imply the same qualitative results considerably weakens the case for rational choice models. For, with this insight in hand, we know that reasons other than the qualitative predictions must be used to discriminate between competing behavioral models. If it comes to finding true rather than merely potential explanations the counterfactual character of the model of fully rational behavior forms a crucial weakness of the purely teleological approaches. As we stated above already, independent reasons speak very loudly against the extreme assumptions of full teleology and, for that matter, also against that of a complete lack thereof. In all likelihood the truth will lie somewhere between the extremes of teleology and evolution.

That less can be explained in terms of rational choice than economists tend to admit seems to be clear. In particular the common knowledge assumptions as well as assumptions of theory absorption mentioned in the introductory remarks are problematic. Evolutionary approaches may avoid such assumptions. By such ways of modeling behavioral dispositions, which are acquired once and forever, can also find their way into economic analysis. We all follow fixed behavioral programs. When the alarm clock rings, we do not always decide anew whether or not to get up. When shopping in a supermarket most of us have once and for all made up their mind to pay and not to steal if the opportunity shows up. To be trustworthy and to fulfill promises is among our "virtues". We tend to be generally virtuous or not, rather than calculating each case on its own merits.

Behavioral dispositions are an important aspect of bounded rationality. Therefore introducing ways of modeling such behavioral dispositions supports the general trend towards theories of bounded rationality in economics. But, of course, not all humans are the same in these

regards. Heterogeneity between individuals may prevail if it comes to rational decision-making and self-management. Some may follow routines where others make forward-looking rational choices in a strategic manner. There will also always be situations in which strategic choice in itself is regarded as appropriate and others in which it is deemed inappropriate. Whether we enter the market to exchange goods or the forum for an exchange of opinions will make a difference at least for many individuals. As theorists we have to consider all these factors and see to it that our models follow suit by locating them appropriately somewhere between the extremes of farsighted teleology and blind evolution. We hope that the present contribution may assist researchers in making such modeling choices.

8. Appendix

(1) Analysis of rest points

(1a) The case $2C > (\underline{\mu} + \overline{\mu} - 1)s(1-s)$.

First note that there can be no rest point (p*, q*, u*) with 0<u*<1:

u* $\in$ (0,1) implies p*=s what, in turn, yields

$$\dot{q}_t = k[(\underline{\mu} + \overline{\mu} - 1)s(1-s) - 2C] < 0\, q$$

and, therefore, q*=0. But then

$$\dot{p}_t = h[u^*(r-1)] < 0$$

and, therefore, p*=0 contradicting p*=s.

Thus, under the above constraint it only remains to explore the border cases u*=0 and u*=1, resp. the cases p*<s or p*>s.

Case u*=0 or p*<s implies

$$sign(\dot{q}) = sign[p^*(\overline{\mu} + (1-\underline{\mu} - \overline{\mu})s) - (1-\underline{\mu})s - 2C].$$

Due to p*<s the right hand side must be compared as follows

$$p^*(\overline{\mu} + (1-\underline{\mu} - \overline{\mu})s) - (1-\underline{\mu})s - 2C] < s(\overline{\mu} + (1-\underline{\mu} - \overline{\mu})s) - (1-\underline{\mu})s - 2C$$
$$= s(1-s)(\mu + \mu - 1) - 2C < 0.$$

This shows that, under the above constraint, the sign of $\dot{q}_t$ is negative so that q*=0. But then p*=0 due to q*=0. This proves (under the constraint) the existence of rest points

(p*, q*, u*) with p*<s, q*=0 and u*=0

to which we refer as "trust dilemma-rest points".

Case u*=1 or p*>s implies

$$sign(\dot{q}) = sign[\, p^*(\bar{\mu} - 1 + (1 - \underline{\mu} - \bar{\mu})s) + \underline{\mu}s - 2C\,].$$

Again the right hand side can be compared in a similar way:

$$p^*(\bar{\mu} - 1 + (1 - \underline{\mu} - \bar{\mu})s + \underline{\mu}\, s - 2C = -\, p^*(1 - \bar{\mu}) - p^*(\underline{\mu} + \bar{\mu} - 1)s + \underline{\mu}\, s - 2C$$
$$< s(1-s)(\underline{\mu} + \bar{\mu} - 1) - 2C < 0$$

due to –p*<-s . Thus $\dot{q}_t$ is negative for p*>s what implies q*=0. Thus it follows from $sign(\dot{p}) = sign(r-1)$

that also $\dot{p}_t$ is negative contradicting that p*>s (>0). This shows that, under the above constraint, only "trust dilemma-rest points" are possible.

(1b) The case $0<2C< (\mu + \mu - 1)s(1-s)$.

We want to prove the possible existence of bimorphic rest points in the sense of (p*,q*,u*) with $s < p^* < 1$, $0 < q^* < 1$ and u*=1. Clearly, p*>s implies $\dot{u}_t > 0$ and thus sooner or later u*=1. Assuming u*=1 such a bimorphism requires $\dot{p}_t(p^*, q^*, 1) = 0 = \dot{q}_t(p^*, q^*, 1)$, i.e.

$$(q^*\bar{\mu} + 1 - q^*)r = q^*(1 - \underline{\mu}) + 1 - q^* \text{ or}$$

$$q^* = \frac{1-r}{\underline{\mu} - (1-\bar{\mu})r} \quad , \quad \text{resp.}$$

$$p^*\mu + [p^*(1-\mu) + (1-p^*)\mu]s - 2C = p^* \quad \text{or}$$

$$p^* = \frac{\underline{\mu}s - 2C}{1 - \bar{\mu} + (\underline{\mu} + \bar{\mu} - 1)s} \ .$$

To guarantee q* ∈ (0,1) thus requires (due to $\bar{\mu}, \underline{\mu} > 1/2$) only q*<1 or

$$r > \frac{1-\underline{\mu}}{\bar{\mu}},$$

whereas p* ∈ (s,1) follows from the case restriction (1b).

(2) Simulation runs

Our simulation of direct evolution is based on discrete time.

Let Δx_t denote $x_{t+1} - x_t$ or $x_{t+1} = x_t + \Delta x_t$ for $x = u, p, q$ so that

$$\Delta u_t = a[p_t - s],$$

$$\Delta p_t = h[\{q_t \bar{\mu} + (1-q_t)u_t\} r - q_t(1-\underline{\mu}) - (1-q_t)u_t]$$

$$\Delta q_t = k[p_t \bar{\mu} + [p_t(1-\bar{\mu}) + (1-p_t)\underline{\mu}]s - 2C - u_t p_t - (1-u_t)s]$$

For s=1/2, r=1/2, $\bar{\mu} = \underline{\mu} = 0.9$, and C=0.005 and given initial values u_0, p_0, q_0 (=0.5) we have simulated the development of u_t, p_t and q_t in periods t=0,1,2,3….. quite systematically by varying coefficients a, h, and k. Here we confine ourselves to illustrating only the most typical runs. We consider "high" resp. "low" values of a, h, and k in order to see the influence of the speed of adjustment on the convergence behavior of the process. Note, that the numerical specification of s, r, $\bar{\mu}, \underline{\mu}$, and C implies that we are in case (1b), i.e. the case of bimorphic rest points.

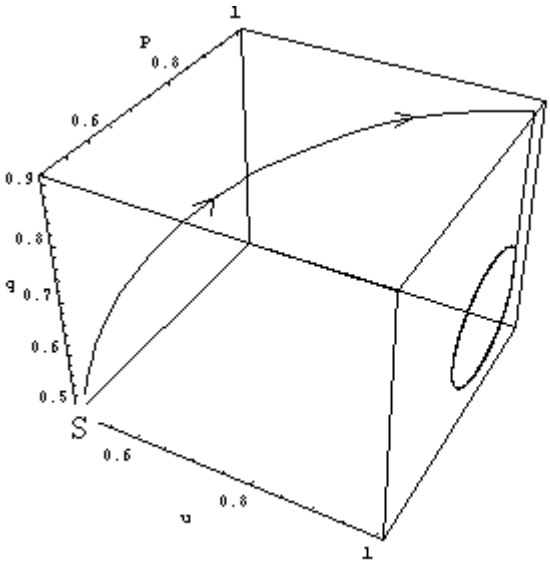
Such a simulation may, of course, lead to values $u_t, p_t, q_t \notin [0,1]$ what can be avoided analytically in continuous time modelling (see Güth, Kliemt and Peleg, 1999) and may be captured in a simulation by simply imposing actual changes $\Delta^* x_t$ instead of Δx_t where

$$\Delta^* x_t = x_t \text{ if } \Delta x_t < -x_t, \text{ and } = 1 - x_t \text{ if } \Delta x_t > 1 - x_t .$$

For the numerical specification we selected several parameter constellations varying the initial values of the structural parameters *a, h, k* of which we present only 3. We denote these constellations by "DES x" with x=1, 2, 3, where DES denotes the abbreviated expression "Direct Evolution Simulation".

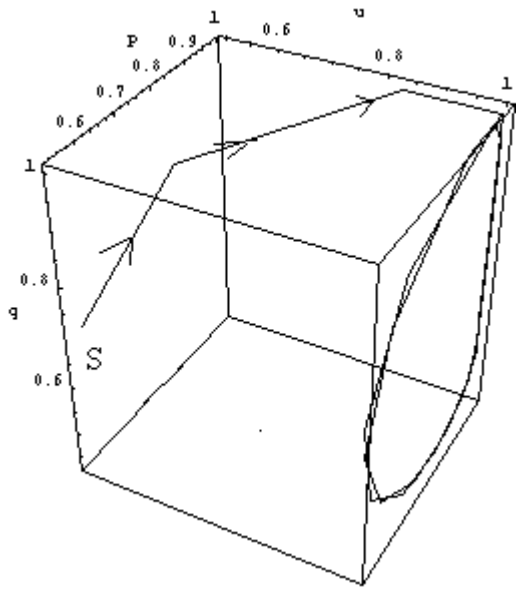
The results of the simulations are illustrated graphically below where the starting point is always indicated by "S" and the movement from S on by direction arrows " $\rightarrow$ " pointing to the future constellations. DES 1 describes a monotonic initial movement to a then stationary cycle around the rest point (p*, q*, u*=1) with 1>p*>s, q* $\in$ (0,1). In case of DES 2 the cycling involves a larger "circle".

DES 1 (low adjustment):



a = 0.1, h = 0.1, k = 0.1

DES 2 (fast adjustment):



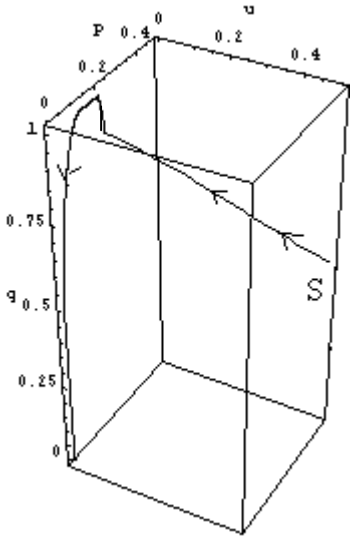
a = 1.1, h = 1.1, k = 1.1

In case (1b) we additionally require the ineqality

$$r > \frac{1-\underline{\mu}}{\overline{\mu}}$$

to hold what is guaranteed by our particular numerical specification above. The relevance of this condition is easy to see by simulation runs which show that the adaptation process shows completely different behavior when the above inequality is violated. Particularly, we set r = 0.1 what implies r=0.1<0.1/0.9. The resulting simulation run is shown below and describes a monotonic convergence to a trust dilemma rest point (p*, q*, u*) with p*<s, q*= 0 =u*.

DES 3:



a = 0.1, h = 0.1, k = 0.1 ; r = 0.1

References

Dacey, Raymond. 1976. "Theory Absorption and the Testability of Economic Theory." *Zeitschrift für Nationalökonomie*, 36:3-4, pp. 247-67.
Frank, R. 1987. "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *The American Economic Review*, 77/4, pp. 593-604.
Güth, Werner and Hartmut Kliemt. 1998. "The indirect evolutionary approach: Bridging the gap between rationality and adaptation." *Rationality and Society*, 10 (3), pp. 377-399.
Güth, Werner and Hartmut Kliemt. 2000. "Evolutionarily Stable Co-operative Commitments." *Theory and Decision*, 49, pp. 197-221.
Güth, Werner, Hartmut Kliemt, and Bezalel Peleg. 1999. "Co-evolution of Preferences and Information in Simple Game of Trust." *German Economic Review*, 1:1, pp. 83-110.
Güth, Werner, and Bezalel Peleg. 2001. "When will payoff maximization survive? - An indirect evolutionary analysis." *Evolutionary Economics*, 11, pp. 479-499.
Morgenstern, Oskar and Gerhard Schwödiauer. 1976. "Competition and Collusion in Bilateral Markets." *Zeitschrift für Nationalökonomie*, 36:3-4, pp. 217-45.
Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium in Extensive Games." *International Journal of Game Theory*, 4, pp. 25-55.
Selten, Reinhard. 1988. "Evolutionary Stability in Extensive Two-Person Games - Correction and Further Development." *Mathematical Social Science*, 16, pp. 223-66.