The Coevolution of Trust and Institutions in Anonymous and Non-anonymous Communities

Werner Güth & Axel Ockenfels

Max Planck Institute for Research into Economic Systems[*]

prepared for:
M.J. Holler, H. Kliemt, D. Schmidtchen and M. Streit (eds.),
Jahrbuch für Neue Politische Ökonomie, 20,
Tübingen: Mohr Siebeck, forthcoming.

March 1, 2002

Abstract

We report on a research program that employs the indirect evolutionary approach to analyze how the institutional environment drives the evolution of trust and trustworthiness through the evolution of moral preferences, and how in turn the evolution of preferences shapes the evolution of the rules of the game. In particular, we describe how the ability to detect trustworthiness in non-anonymous communities supports the evolution of trust and thus crowds out legal institutions. If anonymous interaction prevents type detection, legal institutions such as courts and legal insurance may play a decisive role for the emergence of trust.

---

[*] Max Planck Institute for Research into Economic Systems, Strategic Interaction Unit, Kahlaische Straße 10, D-07745 Jena, Germany; e-mail: gueth@mpiew-jena.mpg.de or ockenfels@mpiew-jena.mpg.de. Both authors gratefully thank Steffen Huck and Roland Kirstein for helpful comments, and gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft.

I. Introduction: The basic trust game

Trust is ubiquitous to almost every economic transaction (see e.g., Arrow, 1974). However, self-interest often dictates not to trust or not to be trustworthy. As a consequence, economic transactions that could make everybody better off are not carried out, unless trust and trustworthiness are reliable (probably evolutionarily or culturally acquired) traits. The trust game in Figure I.1 describes such a social dilemma inherent to many economic scenarios.

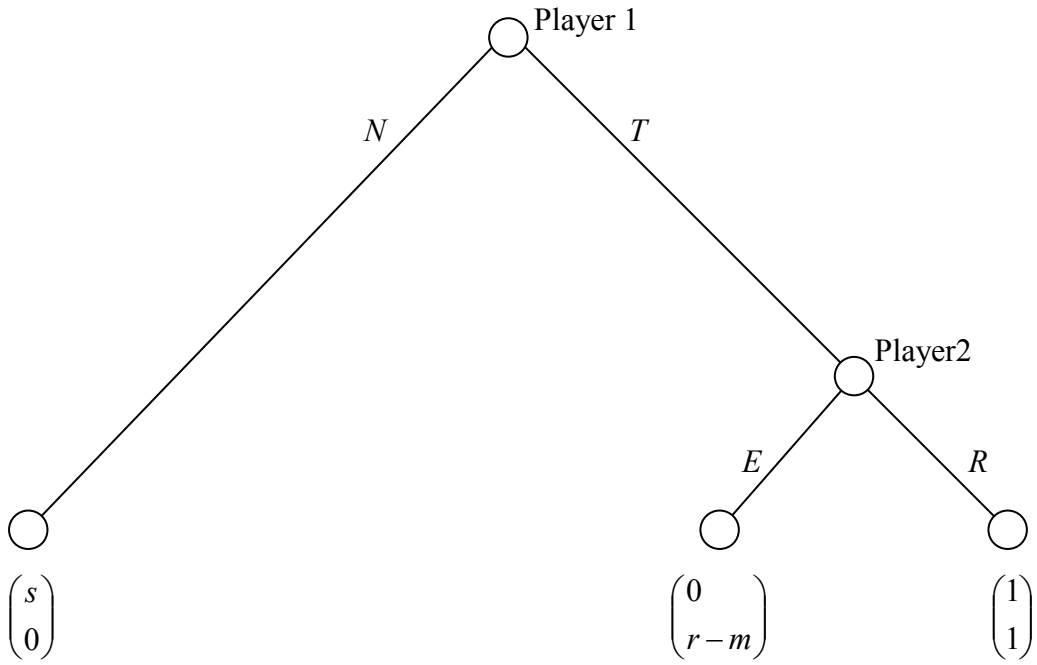

Figure I.1. The basic trust game ($0 < s < 1 < r$)

First player 1 decides between N (*n*on-cooperation) and T (*t*rust in player 2's reciprocity). After N the game ends with payoff s for player 1 and zero for player 2. After T player 2 can either reward (R) or exploit (E). Assume for the moment that the payoff parameter m of player 2 after T and E is zero. Due to $r > 1$ player 2 would exploit and, anticipating this, player 1 would rely on N since $s > 0$. For $m = 0$ common rationality (plus player 1's awareness of 2's rationality) therefore dictates (N, E) and thus a payoff dominated payoff vector since both players would gain by playing (T, R) instead.[1]

[1] The payoffs have been standardized without loss of generality in order to reduce the variety of parameters.

Incentives structures such as illustrated by the trust game (with $m = 0$) frequently occur both in non-anonymous environments where players possess more or less reliable information about the trustworthiness of their opponents, as well as in anonymous communities, where players cannot easily learn their transaction partners' morals. As an example for a non-anonymous scenario, one can think of a bank (player 1) that has to decide whether to trust a local entrepreneur (player 2) by investing in the entrepreneur's project (*T*) or not (*N*). A loan is profitable for the bank if and only if the entrepreneur is trustworthy and thus puts sufficiently effort into the project (*R*), so that the investment (s) plus interest rate ($1 - s$) is eventually paid back. But a payoff-maximizing entrepreneur would prefer to exploit the bank's trust (*E*) and not to pay back anything. As a consequence, the bank should not trust the entrepreneur unless it has reasons to believe that the entrepreneur is trustworthy and refrains from cheating.

In small, non-anonymous communities such as small villages or tribal communities, information about a player's trustworthiness can be conveyed by rumors (Alexander, 1987), by earlier experience with the same customer (Axelrod, 1984), or by other type detection mechanisms such as reliable clues to deceit via physical symptoms of emotional arousal in face-to-face interaction (Frank, 1987; see Ockenfels and Selten, 2000, and Brosig, forthcoming, for mixed experimental results along these lines). Equipped with such clues about the relevant behavioral traits of the local entrepreneurs, a bank may well decide to trust in (some of) its customers.

Things are different for communities with anonymous and infrequent interaction among its members, where signals about others' trustworthiness are not easily available. As an example, think of the kind of sequential transactions that take place on C2C online auction platforms such as eBay (see Ockenfels, 2002a, for a description of the institutional details of eBay's platform). Here, the winning buyer of an auction is typically supposed to send a check or cash to the seller, who, after having received the money, is supposed to complete the transaction by sending the item to the buyer. However, once the seller received the money, he often has no incentives to send the item at all, or he could send only minor quality to the buyer. Also, since repeated interaction with the same opponent is unlikely and since face-to-face communication media are not provided, the information about the opponent is necessarily of poorer quality than the information available in small groups. So why should a seller on eBay be trustworthy and send the item as promised, and, by the same token, why should a buyer trust in the seller and send money? One attempt to promote trust in such an unfavorable environment is to provide buyers with seller-specific information through an

electronic feedback system, such as eBay's 'feedback forum', that imitates the kind of detection mechanisms available in small groups. Another attempt is to employ legal institutions such as court rulings and legal insurance that may, as we will show, support the emergence of trustworthiness even among complete strangers both in online and offline communities.

We examine the evolution of trust and trustworthiness and its dynamic interaction with type detection capabilities and legal institutions with the help of the indirect evolutionary approach. Indirect evolution offers to combine (boundedly) rational decision making, i.e. the traditional approach in economics, with purely adaptative ideas, i.e. the traditional approach in (evolutionary) biology and partly also in psychology. Not only behavior can adapt over time, but also any of its determinants. Indirect evolution therefore offers the chance to endogenously derive what usually is assumed as exogenously given, namely the rules of the game.

II. Non-anonymous communities with type detection capabilities

Trustworthiness can be modeled by 'moral' preferences in the sense that a moral player 2 regrets to exploit a trusting player 1. This can be captured by an immaterial payoff component $m > r - 1$ as shown in the game in Figure I.1, where the interpretation of m is that of intrinsic motivation. In the evolutionary model, the payoff component m is not directly related to reproductive success.

A player 2 with $m > r - 1$ will rationally reciprocate trust by choosing R. A player 2 with $m < r - 1$, on the other hand, will exploit trust by choosing E. In this section we show how the capability to detect the opponent's type m, a measure of an individual's trustworthiness absent of material incentives to be trustworthy, positively affects the evolution of trust and trustworthiness.

Assume a simple evolutionary process in which players from the entire population are matched randomly with equal probability to play one game at each stage. Each player has the same probability of being player 1 or player 2. In each of the interactions, all game parameters including, for the moment, the m-type of player 2 are common knowledge, that is, all players are perfectly informed about their co-player's type. Of course, the m-type of player 1 is irrelevant, but player 2 chooses solely according to his m-type. In such a game of complete information, player 1 will perfectly discriminate between a trustworthy and untrustworthy player 2; he will choose T if and only if he is dealing with a player 2 with type $m > r - 1$. Hence, trustworthy players, when assigned the role of player 2, are trusted and receive a

payoff of 1, while untrustworthy players only receive a payoff of 0 in this case. As a consequence, regardless of how the population is composed of trustworthy and untrustworthy players, the trustworthy fare better when assigned the role of player 2 than the untrustworthy, while both types fare equally when in the player 1 role. All monotone evolutionary dynamics will therefore conclude that a momorphic population composed solely of trustworthy players is a globally stable rest point.

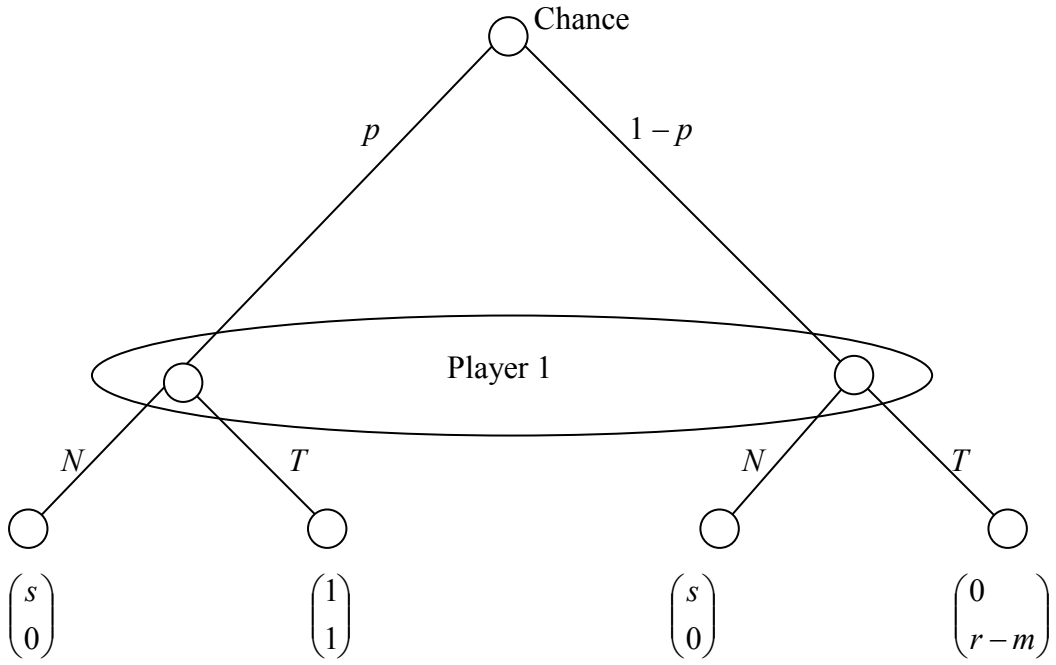

Figure II.1. The (truncated) game of trust when *m* is private information

This optimistic result critically depends on the assumption that perfectly reliable information about all transaction partners' morals is available at no cost. If there is no information about the opponent's type, trust and trustworthiness cannot emerge as the result of evolutionary competition within our model. To see why, assume that the players do not know their co-player's type or, alternatively, assume that such information is prohibitively costly. Let p with $0 \le p \le 1$ denote the population share of individuals whose personal parameter m satisfies $m > r - 1$. Here, as in the following, we assume that p is commonly known and thus determines the beliefs of player 1. Thus with private information about m the game of trust is illustrated by Figure II.1 after truncating the decision of player 2 according to

their *m*-types (recall that *m* with $m > r - 1$ choose *T* and *m* with $m \leq r - 1$ the move *E*).[2] If $p > s$ a rational player 1 will trust (choose *T*) whereas for $p \leq s$ the dilemma result pertains since 1 chooses *N*.

A player assigned the role of player 2 receives a payoff of 1 if trustworthy and a larger (objective) payoff of *r* if untrustworthy, whenever player 1 trusts him. If we assume trembles in the sense of Selten (1983 and 1988), the probability that player 1 chooses *T* is always positive, regardless of *p*, implying that untrustworthy players 2 always fare better. As a consequence, *p*, the population share of trustworthy individuals, vanishes entirely due to the evolutionary pressure. If we apply weaker dynamic stability concepts, there may be also evolutionarily stable strategies in which players 1 never choose to trust their co-players, because the share of trustworthy players is too small. In this case, if trembles are not allowed, there is no evolutionary pressure on the trustworthy to become untrustworthy, just because there is no trust that can be reciprocated or exploited. So a small share of trustworthy may survive evolutionary competition. However, regardless of whether the evolutionary dynamics are driven by trembles or not, we cannot expect to see efficient outcomes in a scenario where reliable information about the co-players' types is not available and where no other institutions may substitute *m*-detection mechanisms.

In a more realistic environment placed in the middle between the two extreme cases of no and perfect information, Güth and Kliemt (forthcoming) have shown that, if some reliable information is available at sufficiently low costs, trust and trustworthiness may still emerge. However, as long the information is not perfectly reliable, the share of trustworthy, *p*, is smaller than 1, implying that such mechanisms generally fail to support complete efficiency as an evolutionarily stable state of affairs. Summing up, the theoretical analysis demonstrates that the better the type detection technology the better are the chances for trust and trustworthiness to evolve (see Bolton et al., 2001, and the references cited therein for supporting experimental evidence along these lines).

It is worth noting, that type detection does not only work in small groups through, say, face-to-face interaction. In fact, large Internet market platforms such as eBay.com or half.com put a lot of effort in imitating small group detection mechanisms via feedback systems that reveal some information about the opponent's type, even when sellers and buyers hardly know more about each other than their user IDs, and when transaction partners rarely meet

[2] The chance move in Figure II.1 is purely fictitious and captures player 1's beliefs concerning player 2's *m*-type which is known to player 2.

repeatedly. In particular, eBay provides information about the opponent's type at no cost via its 'feedback forum' that collects and publicly reveals the assessments by buyers and sellers on each other after a transaction. A person's feedback rating is supposed to "answer many questions about how a person does business", as eBay puts it. That is, in terms of our framework, the feedback rating reveals information about a player's type. eBay's founder Pierre Omidyar describes the merit of eBay's feedback forum as follows: "By creating an open market that encourages honest dealings, I hope to make it easier to conduct business with strangers over the net. Most people are honest. […] But some people are dishonest. […] But here, those people can't hide. We'll drive them away." [3]

Many studies of the trust game with imperfectly reliable signals about the opponent's type (such as Güth and Kliemt, forthcoming) assume that player 1 chooses whether to invest in more or less costly signals about the opponent's type before he decides whether to trust player 2 (for instance by hiring a detective). A related approach is to let player 2 invest in more or less costly efforts to signal trustworthiness in order to attract trusting players 1. This approach is connected to eBay's 'verify program' by which eBay offers their users to establish a proof of identity "so others will trust you as their trading partner." In this program, eBay crosschecks personal information (such as name, date of birth, certain installment and credit accounts and their associated monthly payments) against consumer and business databases for consistency. A user who is successfully verified receives an "ID Verify icon" in his feedback profile that can be seen by all other users. Similarly, sites such as squaretrade.com offer (costly) seals that "increase buyers' trust." Before one receives a seal, there is usually a review process that includes verifying the identity, checking the selling track record and history of resolving disputes, and verifying pricing and return policies. How the opportunity to improve the reliability of one's own signal (rather than to improve the reliability of the opponent's signal) affects the evolution of trust and trustworthiness can obviously be analyzed in similar ways and will also render bimorphic populations as evolutionarily stable if the costs of the seal programs are sufficiently low.

The seal programs as well as eBay's feedback forum can be interpreted as a type detection device to break the anonymity among its users. But they can also be interpreted as a strategic reputation-building device in repeated interactions among strangers (see Kandori's,

[3] Note, however, that there seem to be problems associated with false or biased feedbacks on eBay, implying that the predictive value of the feedback ratings for the opponent's trustworthiness may be severely limited (see Ockenfels, 2002, for a discussion of the merits and the problems of eBay's feedback forum).

1992, and Ockenfels', 2002b, Folk theorem results along these lines). It seems likely to us that seals and feedback ratings reflect both a strategic component and non-strategic type-revelation component.[4] In any case, both interpretations suggest that sellers with a high rating should be more trusted than sellers with a low rating, which actually finds empirical support in field studies based on data from eBay's auction platform (Resnick and Zeckhauser, 2001) and from half.com's fixed-price trading platform (Ockenfels, 2002b).

III. Trust in anonymous communities: The shadow of the court

In many large communities type detection mechanisms are not available and also cannot easily be imitated by rumors or reputations because, for instance, the community members are not (electronically) networked or feedback mechanisms are not implemented. In Güth and Ockenfels (2001) we therefore investigated the role of legal institutions – that are available to both large online and large offline communities – as a mean to promote trust among strangers. To do so, we introduced the trust game with bad luck as shown in Figure III.1.

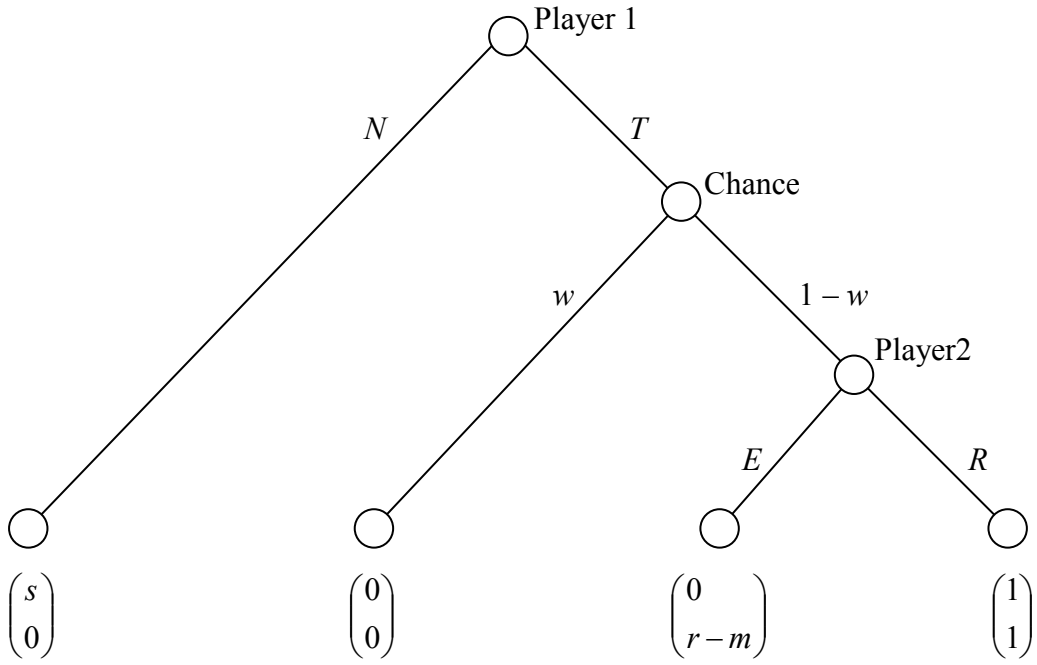

Figure III.1. The game of trust with bad luck ($0 < w < 1/2$)

[4] For non-strategic reputation effects and how they enable mutually profitable cooperation see the experimental study of Albert et al. (2000).

While in Section II we dealt with a scenario in which non-delivery of player 2 implied intended exploitation of trust, we assume here more realistically, that a trusting player 1 may end with a zero payoff for one of two reasons. First, player 2 breaks his promise and chooses *E*, and second, uncontrollable bad luck, which occurs with probability *w* in Figure III.1. That is, a chance move determines whether keeping the promise is impossible, i.e., whether an uncontrollable damage occurs or not. If keeping the promise is impossible, player 1 (the bank or the eBay buyer in our examples in the introduction) loses his investment *s* and both players receive zero payoffs. If compliance is possible, player 2 decides whether to keep the promise, i.e., whether to reward (move *R*) or to exploit (move *E*) the trust of player 1. As before, if the contract is fulfilled, the investment is rewarded by a positive net return of $1 - s$ and player 2 receives a payoff of 1. If player 2 exploits player 1's trust, however, player 1 loses his investment, while player 2 receives a material payoff of *r*.

On eBay, the question of whether a buyer is disappointed because of bad luck (e.g., because the item was lost or damaged by the postmen), or because of fraudulent behavior of the seller (e.g., by an inappropriate item description in the auction) is often at the heart of disputes between buyers and sellers. Therefore, eBay offers various forms of help if a buyer claims that he paid for an item but never received it or received an item that is less than what is described. One attempt to resolve disputes between buyers and sellers is the use of online dispute resolution services such as squaretrade.com recommended by eBay. Squaretrade can be interpreted as an online court that tries to resolve disputes involving "non-delivery of goods or services, misrepresentation, improper selling practices, un-honored guarantees or warranties, unsatisfactory services, credit and billing problems, unfulfilled contracts, etc." However, squaretrade.com does not know the reason for non-delivery better than any other outsider. So one of the central questions is whether (online or offline) courts can support the evolution of cooperation even if the reasons for non-delivery are not observable.[5]

[5] Besides dispute resolution, verified identity program, and the feedback mechanisms, eBay also offers insurance against fraud and recommends the use of (costly) escrow services that allow a buyer to inspect what is bought before sending payment, and the use of (costly) authentication services to get a second opinion on the item by using outside expertise.

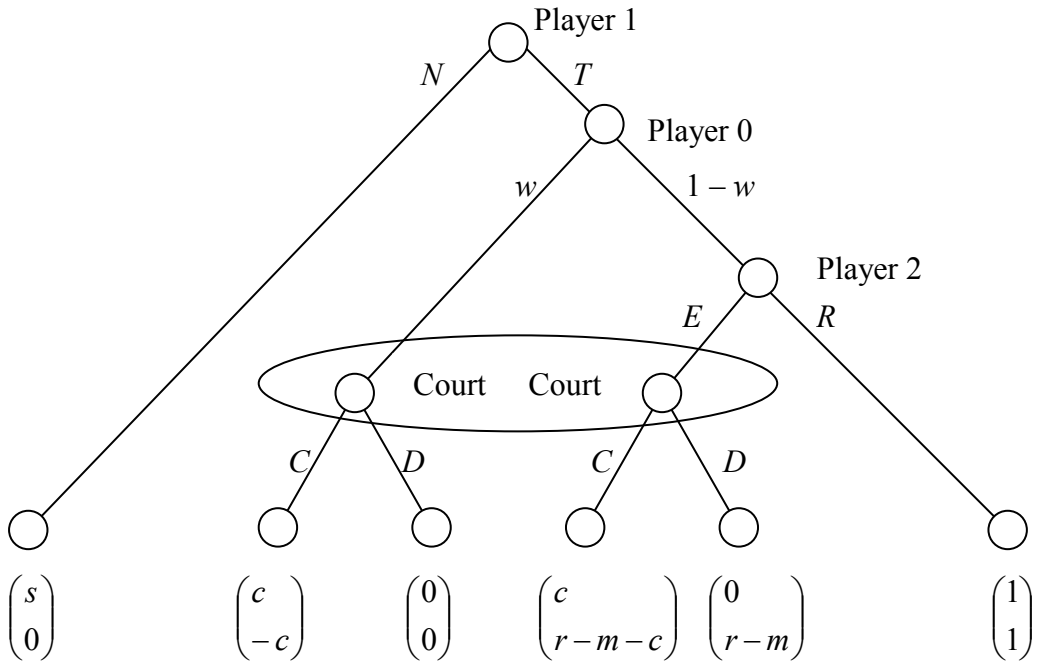

Figure III.2. The court game (Güth and Ockenfels, 2001)

To answer the question how courts can support trustworthy behavior, Güth and Ockenfels (2001) introduced the court game as shown in Figure III.2. (The figure omits for the sake of an easier illustration the fictitious chance move in Figure II.1.) We assume that, whenever player 1 receives a zero-payoff, he appeals to the court. The court then either convicts player 1 (the move *C* in Figure III.2 yielding *c* for player 1) or dismisses the case (the move *D* in Figure III.2).[6] Thereby, the court follows a verdict rule $v = (c, q(e))$, $c \geq 0$, $q(e) \in [0,1]$. This rule determines a probability *q* of conviction and a compensation payment *c* that player 2 has to pay to player 1 in case of conviction. The only indication that can be used by courts in order to come to a verdict is the (conditional) probability $e = e(p,w)$ of untrustworthy second movers given that a loss for player 1 occurred. This probability depends on *p*, as before defined as the share of trustworthy players, and on *w*, the exogenous probability of an unintended damage. The court may use two instruments in order to enforce contracts and norms, the compensation payment *c* and the conviction probability *q*. The compensation payment is assumed to be constant. For instance, in case of conviction, player 1 (the plaintiff) gets half of his demand ($c = 1/2$) or player 2 (the defendant) has to fulfill the contract as if he is fully liable for the loss ($c = 1$). The most natural assumption about the

[6] The online court squaretrade.com, however, cannot force market participants to follow their verdicts.

conviction probability is that the court simply decides for the most probable alternative, i.e. if according to its conditional probability the court considers it to be more likely that non-delivery is caused by 2's choice of *E*, the court decides in favor of player 1 and convicts player 2. Otherwise the court finds player 2 not liable for compensation and dismisses the case. Thus the court enters the interaction as a rational belief forming institution that rules the verdict whose justification is more likely.

The study of Güth and Ockenfels (2001) is a worst-case scenario for the evolutionary emergence of trust and trustworthiness among anonymously interacting players. Neither the players nor the courts are equipped with detection capabilities beyond rational belief forming on the basis of the 'average reputation' p of players 2 and on the unintended damage probability w; all agents involved, including the courts, are therefore complete strangers. Nevertheless, it turns out that courts can still promote trust (without crowding out morals) and thus serve as a substitute for reputation and special detection capabilities that work in small, non-anonymous communities.

In particular, Güth and Ockenfels (2001) show that if the compensation payment is neither 'too small' nor 'too high', the court can positively influence both the share of 'truly' trustworthy players who are internally committed to norms of trustworthiness and the share of players who are inspired to be trustworthy by external material incentives. At the same time, the court can boost participation in efficiency-enhancing transactions. The underlying reason is that if p is small, conviction in case of non-delivery is likely, so that even the untrustworthy players (those with a small m) are induced to be trustworthy (choose R) by the likely conviction. As a consequence, trustworthy behavior spreads out. On the other hand, if p is large, non-delivery is likely to be caused by unintended damage, which reduces the probability of conviction. However, the trustworthy cannot be identified and rewarded by the court so that untrustworthy behavior becomes relatively advantageous in the evolutionary struggle. This obviously can stabilize an interior rest point p^* with $0 < p^* < 1$ and goodies and badies coexisting in the form of an evolutionarily stable bimorphism.

IV. The coevolution of institutions

IV.1 Including legal insurance

While the previous sections demonstrate how information and institutions shape preferences within anonymous and non-anonymous communities, they only examine the coevolution of preferences and just one institutional aspect such as the coevolution of trustworthiness and court rulings in the last section (since the latter adapts to the degree of

morality in society). However, in reality, *many* competing and coevolving institutional aspects are devised and constantly adjusted to promote cooperation. This motivated us (Güth and Ockenfels, 2002) to continue our approach by incorporating another institutional aspect into the court game which can adapt over time, namely the population share of individuals who are legally insured and therefore as plaintiffs have better chances for achieving a verdict.[7]

Clearly, it depends on the court rulings whether or not legal insurance as an institutional aspect will survive. At the same time legal insurance may also affect morality. Thus, we analyze the coevolution of morality (trustworthiness) with court rulings and legal insurance. This allows us to explore not only whether morality will be crowded out or in by institutional aspects of modern societies but also whether one such aspect crowds out or in another such aspect. To our knowledge, this is the first time that such a challenging co-evolutionary model is devised.

If m is commonly known, the trust game with litigation and legal insurance in extensive form is described in Figure IV.1. (Again, Figure IV.1 omits the fictitious chance move shown in Figure II.1 in order to keep the game tree simple.) As before, non-delivery can be caused by nature (the move with probability w) or by exploitation (2's move E). But now player 1 does not automatically involve the court but he has rather the option of yielding (Y), inducing the same outcome as in the basic trust game in Figure I.1, or of appealing (player 1's move A). In the latter case the court becomes active and rules, as before, according to its posterior beliefs by conviction (C) or by dismissing (D) the case.

In the following, we summarize the parameters, some of which have been introduced before:[8]

- w with $0 < w < 1/2$: nature's probability for excluding delivery
- $\delta = \begin{cases} 0 \text{ if player 1 has legal insurance} \\ 1 \text{ if not} \end{cases}$
- c with $1/2 < c < 1$: 2's compensation payment in case of conviction[9]
- s with $0 < s < 1/2$: player 1's outside option payoff
- ε with $0 < \varepsilon < s$: (intrinsic) satisfaction due to suing player 2

---

[7] Of course, our model is still only a partial model; see Kirstein (1999, 2002) and the references cited therein for descriptions and economic analyses of more legal institutions that may interact with the institutions we deal with.

[8] The parameter restrictions partly avoid too complicated case distinctions.

[9] $c > s$ renders trade as profitable for player 1 even when payment by player 2 requires successful litigation. By $c < 1$ we guarantee that player 1 does not prefer successful litigation over due payment.

- C with $0 < C < 1$: legal cost when case is dismissed[10]
- r with $1 + c > r > 1$: exploitation payoff
- m with $m \in \Re$: (intrinsic) regret (for $m > 0$), respectively spite (for $m < 0$) in case of exploiting
- K with $c > K > 0$: the cost of legal insurance (to be subtracted from material success)

[10] The assumption $C < 1$ seems a natural constraint, because unsuccessful litigation should not be more costly than the stakes at hand.

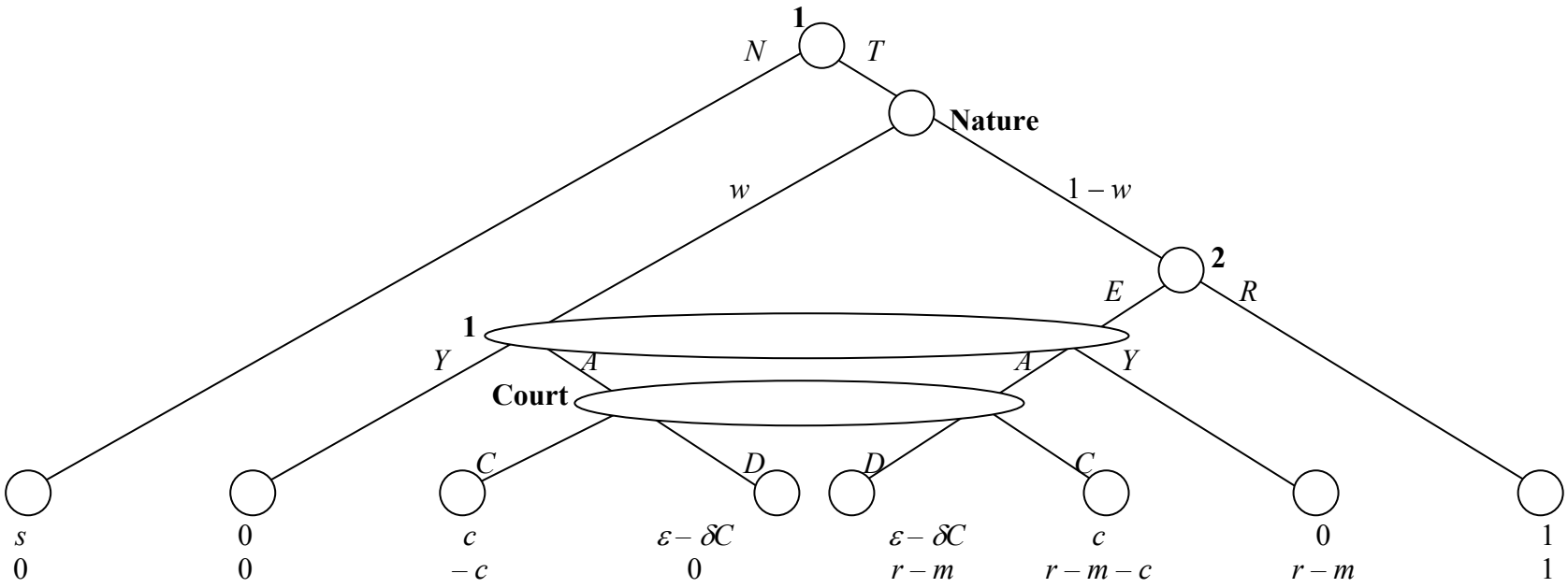

Figure IV.1. The trust game with litigation and legal insurance

The arbitrarily small but positive payoff component ε is interpreted as the satisfaction for not voluntarily yielding to non-delivery. Similar to m, ε is assumed as a given non-material reward. But unlike m, which will be shaped by evolution depending on the relative reproductive success of the possible m-values, ε is merely introduced to solve behavioral ties.

If legally insured ($\delta = 0$), player 1 does not have to pay the legal cost C when the court has dismissed the case. Note, however, that player 1 does not rationally choose for or against legal insurance. Rather, being legally insured or not are assumed to be inherited traits whose diffusion is determined by evolutionary forces. One possible way to interpret the evolutionary diffusion process of legal insurance is to think of it as chosen by contentious individuals who like going to court whenever there is an opportunity to do so. It is well known that such emotional traits, that may not serve immediate material self-interest, may survive evolutionary competition (Frank 1987 and 1988). The difference here is, however, that such an emotional trait also *directly* translates into material payoff consequences caused by the material costs and benefits of the legal institution.

As in Section III, the court plays an ambiguous role in our game model. On the one hand, we do not specify the payoff function of the court player. Actually, we view it as a considerable advantage that we do not burden our approach with more or less arbitrary assumptions concerning the motives of court decision-making. On the other hand, we treat the court player as a rational belief forming institution. The latter is necessary since the predominant principle in civil law is that verdicts hold the party responsible whose guilt is most likely. Thus, courts have to solve the game as any other rational player in order to prepare the ground for applying Bayes-rule when deriving its posterior beliefs. Whatever the solution behavior will be, it is rationally anticipated by the court. This is, as before, common knowledge of the other players as is their rationality. We also emphasize that we will keep our worst-case assumption that courts are complete strangers and do not have any special type detection capabilities.[11]

Since the court rules based on rational posterior beliefs, court decision-making depends on the population composition what further complicates the analysis. Factually, the three institutions, namely

[11] One wonders whether such judicial skills actually exist and, if yes, why such skills are not more widely available. The cost argument (stating that without superior judicial skills the costly legal system would not exist) is, of course, invalid (large and costly cathedrals and temples do not prove the existence of gods and goddesses).

- legal insurance types ($\delta = 0$ and $\delta = 1$)
- trustworthiness (m)
- court ruling (the condition for C or D)

coevolve together. In this paper, we exemplarily demonstrate how such a model of coevolving institutions can be solved. We will restrict ourselves to a special parameter constellation assuming that the players 1 and 2 are non-anonymous, that is we assume that types are commonly known. A thoroughly discussion, including the case of private information, is provided in Güth and Ockenfels (2002).

IV.2 An example

Let z be player 2's probability for choosing R, and y player 1's probability for A. For $y > 0$ the court's posterior probability in Figure IV.1 that player 1 has been exploited by player 2 is well-defined and given by

$$e(z) = \frac{y(1-w)(1-z)}{[w+(1-w)(1-z)]y} = \frac{(1-w)(1-z)}{w+(1-w)(1-z)}.$$

To justify this formula even in case of $y = 0$ we can rely on evolutionarily justified strategy trembles (Selten, 1983, 1988) and the philosophical background of perfect equilibria (Selten, 1975) according to which the unperturbed game should be viewed as a limit of slightly perturbed games for which $y > 0$ is guaranteed.

Since player 2 without legal help or legal advice is hold responsible if E is more likely than unintentional non-delivery, the court will decide for C ($c$onviction) only when $e(z) > 1/2$ or

$$w < \frac{1-z}{2-z} =: f(z).$$

Notice that $f(0) = 1/2$, $f(1) = 0$, and $f'(z) < 0$ for all $z \in [0, 1]$. Thus $f(z)$ can never exceed $w \geq 1/2$ what explains why we restrict the probability for non-delivery without player 2's responsibility to the range $0 < w < 1/2$.

If player 1 is insured and enjoys legal advice, it seems realistic to assume that the odds are improved in his favor.[12] In other words: Player 2 will be convicted even when $e(z) < 1/2$, more specifically, whenever $e(z) > \underline{e}$ with $0 < \underline{e} < 1/2$. With these specific rules governing court decision making for the two types $\delta = 0$ and $\delta = 1$ of the suing party, we can now start to solve the game for all constellations of δ- and *m*-types. With the help of the solution results we then analyze the evolution of the population composition into $\delta = 0$ and $\delta = 1$ types as well as into the various *m*-types for a special case. Note again that since different *m*-types of player 2 might rely on different choice probabilities and since the court rules differently when facing $\delta = 0$- and $\delta = 1$-types, the coevolution of δ- and *m*-types also determines the evolution of court decision making. This justifies our claim that we altogether analyze the coevolution of trustworthiness, legal insurance, and court decision-making.

We assume that types are commonly known, that is, player 1 recognizes player 2's *m*-value, and vice versa player 2 is aware of player 1's δ-type. However, the court only learns the δ-type of player 1 when player 1 appeals and remains completely uninformed about 2's *m*-value. Thus, the court chooses *C* if $e(\hat{z}) > 1/2$ for $\delta = 1$ and $e(\hat{z}) > \underline{e}$ for $\delta = 0$ where $\hat{z}$ denotes the court's expectation concerning *z* based on the true composition $\Pi(\cdot)$ of *m*-types, i.e.

$$\hat{z} = \hat{z}(\Pi) = \int_{m \in \mathfrak{R}} z(m) d\Pi(m),$$

where $z(m)$ is the probability for 2's move *R* by the *m*-type of player 2. Let us consider the following case in which $e(\hat{z}) < \underline{e}$ or

$$w > \frac{(1-\underline{e})(1-\hat{z})}{\underline{e} + (1-\underline{e})(1-\hat{z})}. \qquad (*)$$

In this case, both δ-types would lose when appealing. This, however, will not prevent the $\delta = 0$-type from suing player 2. Thus, in case of (*), only the $\delta = 0$-type will sue (choose *A*) whose population share is denoted by *q* with $0 \le q \le 1$.

When choosing between *E* and *R*, player 2 is supposed to know player 1's δ-type. Since 2's move is, however, independent of $\delta \in \{0, 1\}$, the population share *q* does not matter for 2 in case of (*). Thus, player 2's choice only depends on his own *m*-type as expressed by

---

[12] Such bias can, for instance, result from the fact that experienced lawyers are more likely to avoid obvious form errors such as missing deadlines or not presenting admissible evidence. A more debatable but nevertheless natural justification could be that judges, as many human beings, yield to pressure.

$$z = \begin{cases} 1 \text{ if } m > r-1 \\ 0 \text{ otherwise} \end{cases}.$$

Whenever $m > r - 1$, we will refer to an $\overline{m}$-type of player 2, whereas m-types with $m \leq r - 1$ we summarize as $\underline{m}$-types. As before, all what matters of the m-distribution $\Pi(\cdot)$ is the population share p with $0 \leq p \leq 1$ of $\overline{m}$-types in the population. Consequently, rational expectations of the court imply $p = \hat{z}$ so that (*) becomes

$$w > \frac{(1-\underline{e})(1-p)}{\underline{e}+(1-\underline{e})(1-p)},$$

or equivalently

$$p > \frac{(1-w)(1-\underline{e})-w\underline{e}}{(1-w)(1-\underline{e})}.$$

The right hand side above is positive (recall that by $0 < \underline{e}, w < 1/2$, we have $1 - \underline{e} > w$), and smaller than 1. Thus, our case applies for

$$1 \geq p > \frac{(1-w)(1-\underline{e})-w\underline{e}}{(1-w)(1-\underline{e})}$$

(the other two cases, $\underline{e} < e(\hat{z}) \leq 1/2$ and $e(\hat{z}) > 1/2$, are analyzed in Güth and Ockenfels, 2002).

Let us now derive player 1's initial choice between T and N, denoted by $x := \text{Prob}\{T\}$, for the parameter case solved above. Here, the $\delta = 1$-type in case of T would earn $1 - w$ when $m = \overline{m}$ and 0 otherwise, whereas choosing N yields s, i.e.

$$\delta = 1\text{: } x = \begin{cases} 1 \text{ if } m = \overline{m} \\ 0 \text{ otherwise} \end{cases}.$$

The $\delta = 0$-type would always appeal and then earn ε so that

$$\delta = 0\text{: } x = \begin{cases} 1 \text{ if } m = \overline{m} \\ 0 \text{ otherwise} \end{cases}$$

due to $\varepsilon < s$.

To enter the analysis of the reproductive success of the different types, let us assume for the sake of the usual symmetry of evolutionary games that every individual of an infinite population plays two games, one as player 1 and one as player 2, with randomly chosen partners

in the other role. Reproductive success is given by the solution payoff when setting ε and m equal to 0, i.e. by neglecting the purely intrinsic payoff components, and after subtracting the positive (reproductive success) cost K of legal insurance in case of δ = 0. Our case (*) rules out any conviction and thus $q > 0$ for any stable constellation due to $K > 0$. So only $\overline{m}$ -types will be trusted. Since an m-type earns 1 if $m = \overline{m}$ and 0 otherwise, this implies that p approaches 1 soon or later. For $q = 0$ and $p = 1$, furthermore, the condition of case (1), namely

$$p = 1 > \frac{(1-w)(1-\underline{e}) - w\underline{e}}{(1-w)(1-\underline{e})},$$

is equivalent to $w\underline{e} > 0$ and thus true. This proves that our case contains just one candidate for an evolutionarily stable (q, p)-constellation: (q*, p*) = (0, 1). In fact, as shown in Güth and Ockenfels (2002) this is the only evolutionarily stable population composition when types are commonly known among players. That is, all players are trustworthy and nobody is legally insured. In other words, there is no role for legal institutions if type detection technologies are available (and if courts are not better at type detection than others). On the other hand, if interaction is anonymous, that is if types are private information, there may exist evolutionarily stable rest points in which trustworthy and untrustworthy players coexist and everybody is legally insured, if the costs K of legal insurance are relatively small (Güth and Ockenfels, 2002). Thus, detection capabilities crowd out legal institutions in non-anonymous communities, and legal institutions crowd in trust when detection capabilities are absent. In anonymous communities, courts and legal insurance therefore serve as a substitute of type detection mechanisms.

V. Conclusions

In our evolutionary models, the emergence of trust is either due to detection capabilities in non-anonymous communities or due to legal institutions such as courts and legal insurance when such capabilities are not feasible. In particular, the merits of courts evolve even if judges are not equipped with superior type detection technologies than all other players. Furthermore, our results show that type information renders (costly) legal institutions as useless or crowds them out. So, the decisive aspect of modern interaction may not be its institutions but rather its anonymity that shapes these institutions. Court rulings and legal insurance tend to become crucially important when mankind switched from rural village or tribal life where commonly known types are rather natural, to large, modern communities where type information is often not

feasible or not reliable enough. Unlike large and anonymous metropolitan communities, however, modern online communities that are cheaply connected via electronic communication channels may imitate small group detection technologies through sophisticated computerized feedback systems and thus partly crowd out legal institutions again.

References

Albert, Max, Werner Güth, Erich Kirchler, and Boris Maciejovsky (2000): Exploring Response Behavior - An Ultimatum Experiment. Discussion Paper, Humboldt University, Berlin.

Alexander, R.D. (1987): The Biology of Moral Systems. New York: Aldine De Gruyter.

Axelrod, Robert (1984): The Evolution of Co-operation, New York: Basic Books.

Arrow, Kenneth (1974): The Limits of Organization, New York: Norton, York.

Bolton, Gary, Elena Katok and Axel Ockenfels (2001): What's in a Reputation? Indirect Reciprocity in an Image Scoring Game, Working Paper, University of Magdeburg.

Brosig, Jeannette (forthcoming): Identifying Cooperative Behavior: Some Experimental Results in a Prisoner's Dilemma Game, *Journal of Economic Behavior and Organization*.

Frank, Robert H. (1987): If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?, *American Economic Review*, 77, 593-604.

Frank, Robert H. (1988): Passions Within Reason: The Strategic Role of Emotions, New York: W.W. Norton.

Gauthier, David (1978): Morals by agreement, Oxford: Clarendon Press.

Güth, Werner, and Hartmut Kliemt (forthcoming): Evolutionarily Stable Co-operative Commitments, *Theory and Decision*.

Güth, Werner, and Axel Ockenfels (2000): Evolutionary Norm Enforcement, *Journal of Institutional and Theoretical Economics*, 156(2), 335-347.

Güth, Werner, and Axel Ockenfels (2002): The Coevolution of Morality and Legal Institutions: An indirect evolutionary approach, Working Paper, Max Planck Institute for Research into Economic Systems, Jena.

Kandori, M. (1992): Social Norms and Community Enforcement. *Review of Economic Studies*, 59, 63-80.

Kirstein, Roland (1999): Imperfekte Gerichte und Vertragstreue. Eine ökonomische Theorie richterlicher Entscheidungen. Gabler: Wiesbaden.

Kirstein, Roland (2002): Comment on 'The Coevolution of Trust and Institutions in Anonymous and Non-anonymous Communities.' University of Saarbrücken, mimeo.

Ockenfels, Axel (2002a): New Institutional Structures on the Internet: The Economic Design of Online Auctions, Working Paper, University of Magdeburg.

Ockenfels, Axel (2002b): Reputationsmechanismen auf Internet-Marktplattformen, Working Paper, Max Planck Institute for Research into Economic Systems, Jena.

Ockenfels, Axel, and Reinhard Selten (2000): An Experiment on the Hypothesis of Involuntary Truth-Signaling in Bargaining, *Games and Economic Behavior*, 33(1), 90-116.

Resnick, P. and Zeckhauser, R. (2001): Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. Working Paper, NBER Workshop.

Selten, Reinhard (1975): Re-examination of the Perfectness Concept for Equilibrium in Extensive Games, *International Journal of Game* Theory, 4, 25-55.

Selten, Reinhard (1983): Evolutionary Stability in Extensive Two-person Games, *Mathematical Social Sciences*, 5, 269-363.

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