On the Co-evolution of Retribution and Trustworthiness: An (Indirect) Evolutionary and Experimental Analysis

by

WERNER GÜTH, HARTMUT KLIEMT,

M. VITTORIA LEVATI AND GEORG VON WANGENHEIM

Standard economic explanations of good conduct in trade rely almost exclusively on future-directed extrinsic motivations induced by material incentives. But intrinsic motives to behave trustworthy and to punish untrustworthiness do support trade. In our model, intrinsically motivated players are aware of their own type and observe the population share of other types. The material success of various types and their co-evolution are analyzed, and it is checked whether the dynamics of the indirect evolutionary analysis are replicated in the laboratory. (JEL: B 52, C 72, C 90)

1 Introduction

Folk theorem (AUMANN [1981]) and type uncertainty logic (KREPS, MILGROM, ROBERTS AND WILSON [1982]; KREPS AND WILSON [1982]) suggest that improving an individual's own future prospects can explain punishment behavior. But theoretical philosophical arguments (e.g., MACKIE [1982]), field studies (e.g., WESTERMARCK [1906]) as well as laboratory experiments (e.g., FEHR AND GÄCHTER [2000], ANDREONI, HARBAUGH AND VESTERLUND [2003], MASCLET, NOUSSAIR, TUCKER AND VILLEVAL [2003]) show that often a punishing individual must be intrinsically motivated to go against her own extrinsic or material interests. The costs of such "consumptive" acts cannot be explained exclusively in terms of strategic future-directed choices.[1]

[1] GÜTH, LEININGER AND STEPHAN [1991] show that also accounting for the Smithian discipline of continuous dealings is impossible if no retributive – backward-looking – element is allowed.

We intend to study how retributive responses and (un)fair dealings in trade can influence each other. To this end, we combine a trust game, in which a first mover has to be trustworthy toward a second mover, with a punishment game, in which the trustor can punish a deviant trustee at a cost.[2] We scrutinize to what extent first-movers' non-strategic trustworthiness and second movers' non-strategic proclivity to punish deviant behavior can support each other in repeat interactions. In discussing the issue, we pursue two lines of argument. First, in an indirect evolutionary analysis of the co-evolution of (un)trustworthiness and retributive inclinations, we show that, theoretically, intrinsically motivated punishing behavior should be expected to vanish. Then, in an experimental study of the dynamics of (un)kind behavior and retribution, we investigate whether there are initial conditions under which intrinsic motivations to behave trustworthy and intrinsic motivations to punish the non-trustworthy support each other.

If legal (institutional) design is meant to apply to a "long" future, it should rely on evolutionarily stable intrinsic motivations. The indirect evolutionary analysis warns us not to base legal design on intrinsic motivations since, given our assumptions, they stand to be eroded. But a different message stems from our experimental study, which reveals stable population shares of moral behavior rather than their erosion. This demonstrates that trustworthiness and retribution, unlike other behavioral dispositions, may be rather stable intrinsic inclinations on which we can rely when designing legal institutions.

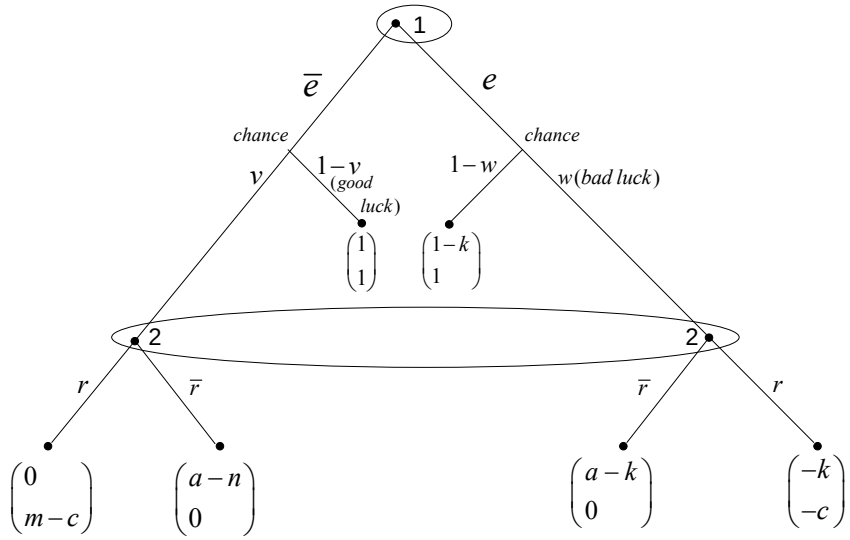
2 The Model

Consider the game tree presented in Figure 1. In interpreting the tree, we think in terms of a stylized story of failure and success in executing a promise or a contract. Failure may be due either to the trustee's negligence or to bad luck. If strict liability applies, the reason behind the failure should not trigger a different reaction of a

[2] We acknowledge Bernd Lahno's remark that interpreting the interaction we present in the next section as a trust game is questionable as we disregard the trustor's decision to contract with the trustee. In our setting, trust is somehow taken for granted and only trustworthiness is modeled.

Figure 1

The Game Tree



potentially punishing individual. But, moral and legal sanctions normally distinguish whether the promisor is a causative factor in the failure or not.

Assume that player 1 has promised to render a service to player 2. The act of promising is not modeled here. The game starts after the promise, with the decision of player 1 to invest proper effort, e, or to shirk, $\bar{e}$. Delivery may succeed or fail depending on "luck". Let us start with the two cases in which delivery succeeds.

After $\bar{e}$, a chance move decides whether, with probability $v \in (1/2, 1]$, player 1's shirking yields a failure to deliver or whether, with complementary probability, "good luck" leads to a proper delivery anyway. In the latter case, the transaction is completed as intended and, under appropriate normalization, each individual gets a payoff of "1".

If player 1 puts in proper effort, e, and if with probability $1-w$, $w \in [0, 1/2)$, no forces beyond her control prevent the occurrence of delivery, the contract is fulfilled in the expected way. In this case, as effort is costly, player 1 receives "1 – k", with $k \in (0,\ 1]$ being the cost of effort, while player 2 gets a payoff of "1".

Next, we consider the two cases in which delivery fails, and the implied rational

beliefs. Let $x \in [0,1]$ be the population share of types who choose "no-effort" when assigned to player 1's role. Then, the *a priori* probability that a shirking individual will cause non-delivery is vx. We assume that this case of failure is the only situation in which the punishing individual intends to punish. After observing non-delivery, player 2's posterior beliefs concerning bad luck and bad intentions of player 1 must be derived from her prior beliefs.[3] In the information set where she decides between revenge r and no revenge $\bar{r}$, we get the probabilities $\frac{(1-x)w}{xv+(1-x)w}$ that failure was due to bad luck, and $\frac{xv}{xv+(1-x)w}$ that the choice of $\bar{e}$ was co-responsible for failure.

We assume that while population shares of choice-types are commonly known, an individual's own type is private information. The implicit infinite population assumption of our theoretical analysis rules out that individuals can use their private type information to condition their beliefs. In the laboratory, individuals could in principle condition their beliefs on their own type (for the extreme case of merely three individuals see GÜTH, GÜTH, AND KLIEMT [2002]). But our experimental population involves 16 participants. This number should be perceived as sufficiently large by real decision makers to ignore their own type.

After the transaction fails, the costs of "revenge" for player 2 are $c(>0)$. Should she decide on punishing player 1 by choosing *r*, player 2 receives material payoffs of "0 – *c*" while player 1 is left either empty-handed, i.e. with "0", if she had put in no effort, or with effort costs "$-k$". Note that we are not dealing here with punishment in a criminal law sense, but rather with non-delivery of a service, which results in restitution regardless of whether there is any guilt involved. In case of no revenge $\bar{r}$ after non-delivery, player 2 receives 0-payoffs while player 1 obtains material payoffs of $a-k$ or a, with $0<a<1$, depending on whether she has exerted effort or not.

[3] A non-stochastic variant of ultimatum bargaining where punishment has to be based on beliefs is the "Yes or No-game" (cf., GÜTH, LEVATI, OCKENFELS AND WEILAND [2005]). In this game, responders do not know the proposal when deciding between acceptance and rejection.

This leaves only the payoff parameters m and n unexplained. They are assumed to represent exclusively intrinsic concerns with no direct effect on substantive or material success. For instance, if two individuals with different *m*-parameters show the same behavior, objectively they may have the same success, but they may be evaluating their actions differently. Except for *m* and *n*, all the other parameters, including "0", serve a twin purpose: they indicate subjective evaluations and objective success at the same time (since actors subjectively evaluate results according to objective success). Only *m* and *n* do not indicate substantive success; they represent intrinsic motivations, and affect substantive success merely *indirectly* through behavior. This explains why we refer to our approach as an *indirect* evolutionary one.

The parameter $m\,(\geq 0)$ represents the influence of purely subjective evaluations like retributive emotions (e.g., a desire for revenge) on preferences. A rational player 2 will use r when it yields a higher expected payoff as compared to $\bar{r}$. This occurs whenever her retributive emotions are strong enough, i.e., satisfy $m > \bar{m}(x) := c\left(\frac{(1-x)w}{x\nu}+1\right)$. On the other hand, she will choose $\bar{r}$ if $0 \leq m \leq \bar{m}(x)$. Thus, it depends partly on the population behavior in the role of player 1 whether a certain *m*-type will engage in revenge or not. Only player 2 is aware of his own *m*-type. We denote by y the population share of players 2 with $m \geq \bar{m}(x)$ where, of course, $y \in [0,1]$.

Similarly, $n(\geq 0)$ stands for intrinsic feelings of uneasiness by player 1 if delivery fails due to her choice of $\bar{e}$, and she remains unpunished. Here, intrinsic motivations of sufficient strength render player 1's choice of e better than $\bar{e}$ if $1-w+w(1-y)a-k>1-v+v(1-y)(a-n)$ or $n > \bar{n}(y) := \frac{v-w}{v}a - \frac{v-w}{(1-y)v} + \frac{k}{(1-y)v}$ for $0 \leq y \leq \underline{y}$, and $n \geq \bar{n}(y) = 0$ for $y > \underline{y}$ where $\underline{y} = \frac{k}{(\nu-w)a} - \frac{1-a}{a}$ and $k/(\nu-w)a > 0$ due to $w < 1/2 < \nu$. In the remainder of the paper, we require $k \in \big((1-a)(\nu-w), \nu-w\big)$ in order to concentrate on the interesting cases. The lower

bound ensures that, at least without any punishing behavior ($y = 0$) and absent intrinsic motivation to exert proper effort ($n = 0$), there is some temptation to shirk. The upper bound makes sure that all players 1 will try to deliver properly when all players 2 have sufficient intrinsic motivation to always punish ($y = 1$). Note that the requirement implies that $\underline{y} \in (0,1)$, and that the larger the population share y of players 2 who are inclined towards revenge, the fewer the players 1 who shirk.

In this way, for all constellations $(x, y) \in [0,1]^2$, we can determine the rational behavior and, thus, the material success of all inclination types. This material success is the solution payoff with m and n being equal to 0 because intrinsic concerns influence reproductive success only indirectly via rational choice-making.

3 The Evolutionary Dynamics

Let us assume that individuals, chosen from an infinite population, play the game of Figure 1 being randomly paired. The beliefs guiding behavior in period t are determined by the commonly known population shares of behavior in period $t - 1$, thereby requiring an assumption about initial conditions (x_0, y_0). Assuming various initial conditions, we can explore to what extent the resulting evolutionarily stable behavioral constellation is path dependent.

For beliefs determined by (x_{t-1}, y_{t-1}), the analysis in Section 2 implies the following average material or reproductive success of the two n-types:

$$R_t\left(n > \bar{n}_t\left(y_{t-1}\right)\right) = 1 - w + w\left(1 - y_{t-1}\right)a - k$$

$$R_t(n < \bar{n}_t(y_{t-1})) = 1 - \nu + \nu(1 - y_{t-1})a\,.$$

The trustworthy types (i.e., those with $n > \bar{n}_t\left(y_{t-1}\right)$) are more successful, and should therefore increase their population share if $(\nu - w)\left[1 - \left(1 - y_{t-1}\right)a\right] > k$ or $y_{t-1} > \underline{y}$. For a population share y_{t-1} of individuals intrinsically motivated to punish (given x_{t-2}), the relative frequency x_t should, if possible, increase as compared to x_{t-1} if

$y_{t-1} \le \underline{y}$, and remain equal to $x_{t-1} = 0$ if $y_{t-1} > \underline{y}$. In the latter case, one has $n \ge \bar{n}_t(y_{t-1}) = 0$ for any *n*, which implies that all players 1 invest proper effort and hence receive the same material payoff, thereby excluding any evolutionary forces on the distribution of *n*.[4]

Similarly, to determine the evolution of y_t, we use the material success of the two *m*-types

$R_t\left(m > \bar{m}(x_{t-1})\right) = -c\left[vx_{t-1} + w(1-x_{t-1})\right] + (1-v)x_{t-1} + (1-w)(1-x_{t-1})$, and

$R_t\left(m < \bar{m}(x_{t-1})\right) = (1-v)x_{t-1} + (1-w)(1-x_{t-1})$.

Clearly, regardless of players 2's beliefs, $R_t\left(m < \bar{m}(x_{t-1})\right) > R_t\left(m > \bar{m}(x_{t-1})\right)$ applies and, consequently, $y_t < y_{t-1}$ when y_{t-1} is not yet minimal. Hence, independently of the size of the population (finite or infinite) and of the consequent type-dependent beliefs, the proportion of individuals with sufficient intrinsic motivations to punish (for given x_{t-1}) will shrink. As to the evolutionarily stable state, we can therefore conclude that $y_t \to y^* = 0$ for $t \to \infty$. But this, in turn, implies that sooner or later y_{t-1} ends up in the range $\left[0, \underline{y}\right)$. The population share of *n*-types with $n \le \bar{n}_t(y_{t-1})$ will increase and thus $x_t \to x^* = 1$ for $t \to \infty$. In sum,

PROPOSITION *The only evolutionarily stable constellation is* $(x^*, y^*) = (1, 0)$, *i.e., a population without trustworthiness and without punishment where first movers (players 1) shirk (choose* $\bar{e}$*) and second movers (players 2) yield (choose* $\bar{r}$*).*

4 Experimental Protocol

In our experimental study, we employ finite populations with 16 individuals who are randomly paired to play the game described in Figure 1. The way the game is explained to participants is in terms of an expert *E* (player 1) who must advise an

[4] The reader should keep in mind that, strictly speaking, evolution takes place in terms of the

investor *I* (player 2) on an investment.[5] In doing so, *E* can exert effort or not, where effort costs are k = 20 ECU.[6] If *E* does not exert effort, the probability that the investment fails is ν = 2/3. Otherwise, this probability is reduced to w = 1/3. In each period, member *E* is endowed with 50 ECU (Experimental Currency Unit), which she can double if the investment succeeds (yielding 100 ECU). In the latter case, investor *I* receives 100 ECU as well. Should the investment fail, member *I* can sue member *E*, where suing costs are c = 20 ECU.[7] If *I* sues, she pays the 20 ECU whereas *E* loses her initial endowment (resulting in 0 ECU); otherwise, *I* pays (and gets) nothing and *E* keeps her endowment (i.e., 50 ECU). Hence, the experimental material payoffs can be 100, 80, 50 – n, 30, 0 or –20 for member *E*, and 100, 0, m – 20 or –20 for member *I*. According to Figure 1, investors should come into play only in case of bad luck (here, investment failure). However, to get richer data on investors' behavior, we asked them to make their suing decision before knowing if the investment succeeded or not.[8]

By monetary equivalents we induce appropriate intrinsic concerns according to which $m > \overline{m}$ and $n > \overline{n}$. In particular, there are two basic treatments, which differ only with respect to the number of individuals who are allocated such equivalents. In one treatment (henceforth, *L*-treatment), only one subject per player role is endowed with monetarily induced "additional" concerns. In the other treatment (henceforth, *H*-treatment), three subjects per player role are endowed with monetarily induced "additional" concerns.[9]

In the *L*-treatment, due to our population's size, the share of players 1/members *E* who should choose no effort is x = 7/8, and the share of players 2/members *I* who should opt for suing is y = 1/8 = 0.125. Given these shares and the other parameter

---

distributions of n and m and not in terms of x and y.

[5] For more details see the instructions reported in the Appendix.

[6] All amounts in the experiment were denoted in ECU, where 10 ECU equated 1 Euro.

[7] Since suing is the only way for *I* to take revenge, behaviorally suing corresponds to revenge.

[8] SELTEN [1967, 1998] provides details about the relevance of the strategy method in experiments. BRANDTS AND CHARNESS [2000] show that this 'cold' methodology does not trigger different behavior as compared to 'hot' play.

[9] We cannot exclude that, due to imported intrinsic moral preferences by participants, the actual shirking and suing ratios will differ from those artificially induced. However, we are mainly interested in investigating the effect of an increase in "artificially induced moral behavior".

values, we obtain $\overline{m}(x) = 21.43$ and $\underline{y} = 0.2$. Hence, in this treatment, $y < \underline{y}$ holds, implying $\overline{n}(y) = 15/7 \approx 2.14$. Appropriate intrinsic concerns are therefore brought about by setting $n > 2.14$ for one subject in the role of player 1 and $m > 21.43$ for one subject in the role of player 2. In particular, we chose $n = 67$ and $m = 30$ in order to have large differences $n - \overline{n}(y)$ and $m - \overline{m}(x)$. The remaining individuals in the population are left to their own resources, i.e., for them $n = m = 0$ holds.

In the *H*-treatment, one has $x = 5/8$ and $y = 3/8 = 0.375 > \underline{y} = 0.2$, implying $\overline{m}(x) = 26$ and $\overline{n}(y) = 0$, respectively. As $y > \underline{y}$ also entails that expected payoffs from exerting effort exceed expected payoffs from shirking for all $n \geq 0$, we deemed it reasonable to induce a rather small amount of trustworthiness and set $n = 9$ for three subjects in player 1's role. Concerning players 2, we chose roughly the same difference $m - \overline{m}(x)$ as in the *L*-treatment and relied on $m = 35$. The other subjects have $m = n = 0$.

Pairs interact for a total of 100 periods in a stranger design (i.e., groups are randomly assembled every round). Monetary equivalents are allocated to the designed number of players at the beginning of the experiment and kept throughout. Subjects can switch role from one period to the next.[10] This implies that, but for period 1, the actual number of monetarily induced subjects may vary in the course of the experiment. In neither treatment, participants are informed about the number of individuals who are induced by specific monetary incentives to pursue an objective function other than the original monetary payoffs. But, in each period, except for the first one, participants receive information about the shares of no effort- and revenge-choices in the previous period.

An experimental session consists of two subsequent phases of 50 periods each. Each phase employs either the *L*- or the *H*-treatment (within-subjects factor) with the order of treatments as between-subjects factor: in half of the sessions subjects

[10] We readily accept the criticism by Theodore Eisenberg that assigning experimental roles with no entitlement may change behavior (see GÜTH AND KLIEMT [2004] for a critical discussion of experiments with "manna from heaven"-rewards). We decided, however, not to provide entitlement

experience the *L*-treatment in the first 50 periods and the *H*-treatment in the last 50 periods while in the remaining sessions they experience the treatments in the reverse order.

5 Experimental Results

The computerized experiment was conducted at the experimental laboratory of the Max Planck Institute in Jena (Germany) in June 2006. The experiment was programmed and performed with the z-Tree software (FISCHBACHER [1999]). Participants were undergraduate students from different disciplines at the University of Jena.

Overall, we ran six sessions with 32 participants each. In each session, we distinguished two matching groups of 16 players, guaranteeing six independent observations for each order (*LH* vs. *HL*). Each session took about two hours. The average earnings per subject were about €23 (including a show-up fee of €2.50).[11]

First, we focus on the behavior of the monetarily non-induced subjects. Figures 2 and 3 display the time paths of the average shares of monetarily non-induced players 1 who did not exert effort (Figure 2) and of monetarily non-induced players 2 who engaged in revenge (Figure 3), separately for the two treatments and for both orders. The main prediction of the indirect evolutionary analysis is clearly rejected. On average, players 1 who are left to their own resources, independently of treatment and treatments' order, are largely trustworthy, and trustworthiness does not vanish. The overall average shares of monetarily non-induced players 1 who do not exert effort in the *H*-treatment are 0.197 under the *HL*-order and 0.207 under the *LH*-order. The respective shares for the *L*-treatment are 0.273 and 0.241. Given this low level of untrustworthiness, it is not surprising that the shares of monetarily non-induced players 2 who punish are rather low throughout each session. In particular, the overall average shares of $m = 0$-players 2 engaging in revenge are 0.063 (0.065) and 0.056

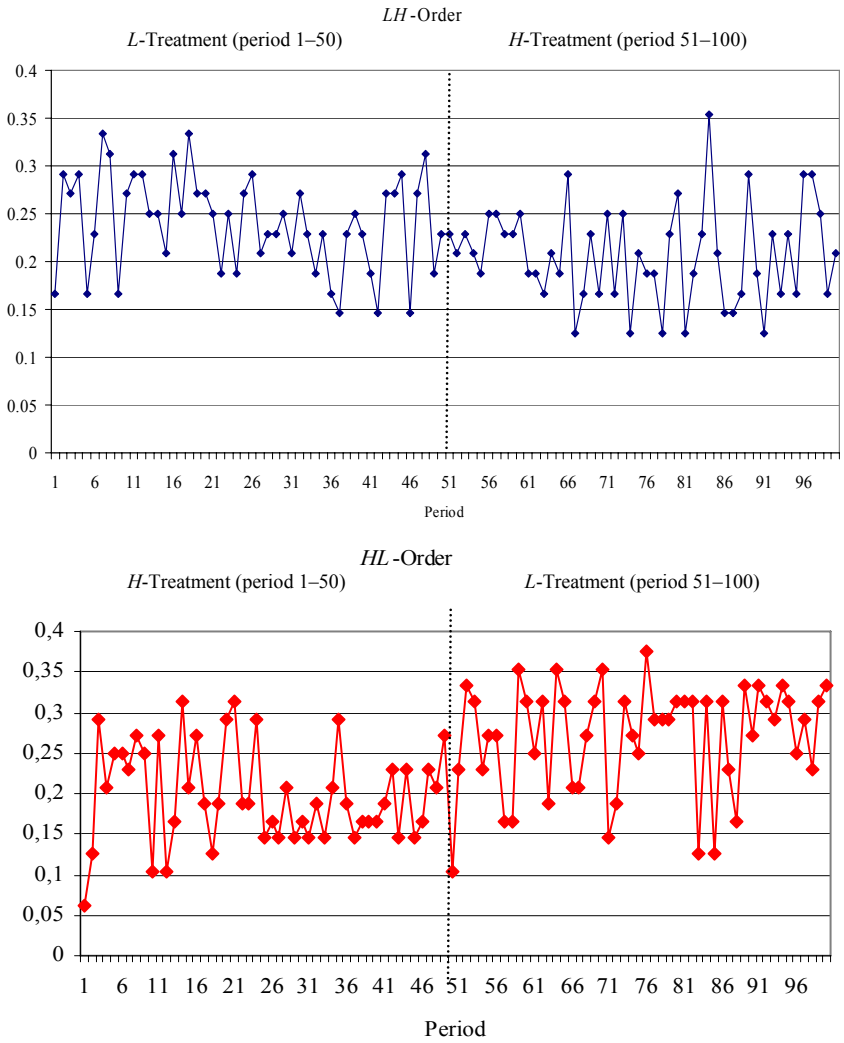
---

because this would have required a much more complex protocol.

[11] In order to avoid portfolio-diversification effects (see MARKOWITZ [1952]), in each session four periods were randomly selected for payment.

Figure 2

Non-Monetarily Induced Players 1's Average Levels of No Effort Over Periods



(0.081) in the *H*- (*L*-) treatment under the *HL*- and *LH*-order, respectively. All these amounts are, nevertheless, significantly larger than zero according to binomial tests (1 percent level).[12]

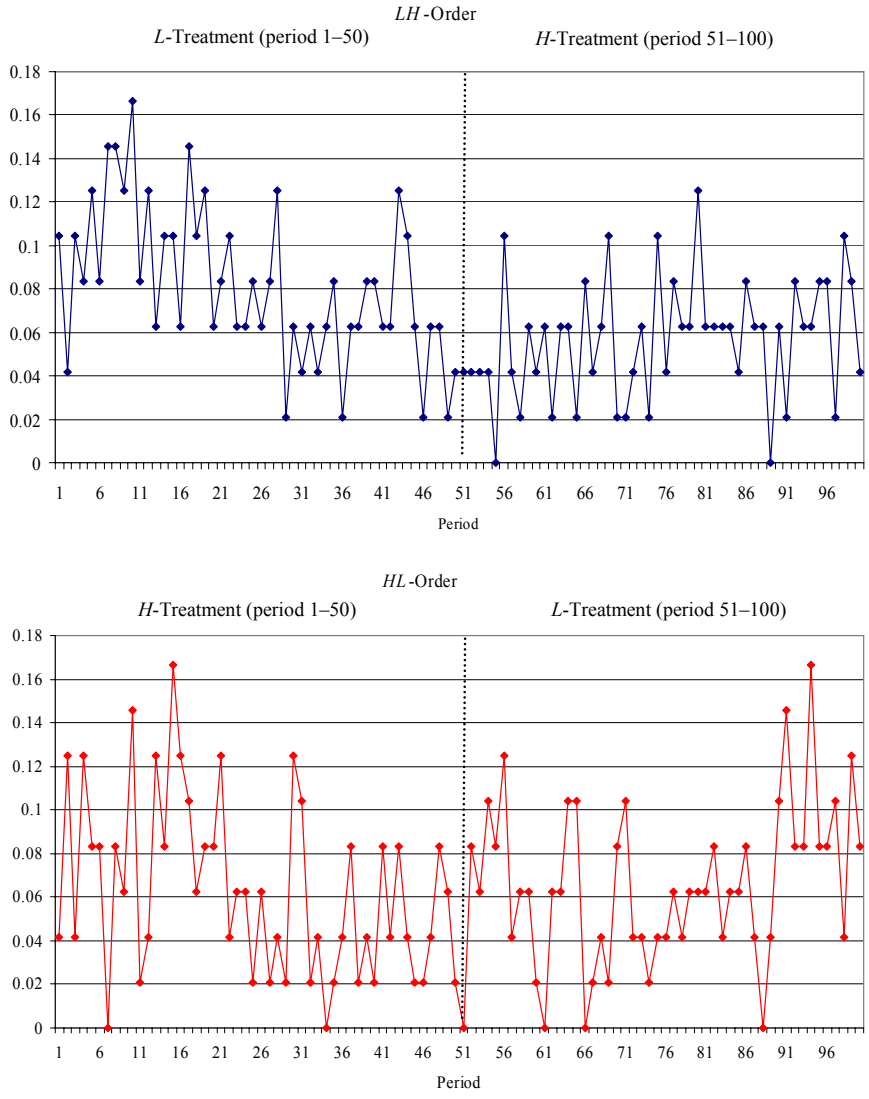
OBSERVATION 1 *On average, monetarily non-induced participants behave trustworthy, i.e., exert effort, and there is a positive share of non incentivized punishment*.

This is in line with previous studies indicating that we can rely on strong intrinsic

[12] The binomial tests rely on the averages over players separately for each period and matching

Figure 3

Non-Monetarily Induced Players 2's Average Levels of Revenge over Periods



motivations allowing for a good deal of cooperation without contract enforcement.

A thorough look at the individual data reveals that about 45% (35%) of "non-induced" participants in the expert-role display moral behavior throughout the experiment in the *H*- (*L*-) treatment under both orders, and further 20% (25%) choose to exert effort at least 75% of the times. A negligible percentage (about 5% in each treatment under either order) of players 1 never exerts effort. As to participants in the investor-role, only one in each treatment and order is found to retaliate over all

---

group.

periods. The percentage of those who sue at least half of the times is around 5%. Note, however, that the moral acceptability of punishing is precarious as bad luck occurs with positive probability even in case of effort, and thus punishment is destined to hit those who do not deserve it. Since experts are mainly nice, systematic threats to punish could drive out trustworthiness.

The choices of monetarily non-induced players in a given treatment do not seem to depend on when the treatment is faced (i.e., in the first or in the second 50 periods). Wilcoxon rank-sum tests (two-sided) comparing independent observations averaged over players for each matching group[13] confirm that, in either treatment, the shares of no effort by players 1 with $n = 0$ are not affected by the order in which the treatment is played ($p > 0.589$ for both comparisons). The same holds when we compare average shares of revenge by players 2 with $m = 0$ in the same treatment under the two different orders (both $p > 0.630$).

OBSERVATION 2 *There is no obvious order effect, i.e., intrinsically motivated participants do not behave differently depending on when they face a given share of monetarily induced morality*.

Next, we compare behavior in the *H*-treatment (where we induce intrinsic concerns for three subjects per player role) and the *L*-treatment (where intrinsic concerns are induced for only one subject per role). Contrary to expectations, Figures 2 and 3 suggest no difference between treatments regarding the choices by non-induced participants. Non-parametric tests confirm that the prompted initial conditions affect neither trustworthiness nor punishment in a significant way.[14]

OBSERVATION 3 *Neither trustworthy nor revengeful actions need to be supported by a high share of incentivized moral behavior*.

Nonetheless, the overall average shares reported above reveal that monetarily non-induced players 1 choose "no-effort" more often in the *L*-treatment than in the *H*-

[13] Unless otherwise stated, all statistical tests are based on statistically independent matching groups.

[14] The lack of statistically significant difference between treatments is detected both when we consider data separately for the two orders and when (based on previous results) we pool data across sessions.

Table 1
Spearman and Kendall Correlation Coefficients between Frequency of Revenge and Observed Share of No Effort

Order	Treatment	Spearman ρ	Kendall τ
HL	*H*	0.116 (0.445)	0.093 (0.396)
	L	0.105 (0.233)	0.082 (0.226)
LH	*H*	0.26 (0.065)	0.183 (0.050)
	L	0.22 (0.062)	0.189 (0.043)

Note: p-values in parentheses

treatment (0.257 vs. 0.202, pooling averages across sessions). At odds with our prediction based on the simple counting of monetarily induced players 2, the frequency of revenge is higher in the *L*-treatment as compared to the *H*-treatment (0.073 vs. 0.060, overall sessions). This indicates that the intentions to punish in a given period are resulting from latent dispositions which trigger action when observing a high share of $\bar{e}$ in the previous period. To test whether there is correlation between inclination to act in revenge and observed untrustworthy behavior, we calculated both Kendall's τ and Spearman's ρ coefficients. Both methods show that, whatever the treatment and the order, this correlation is positive, but (weakly) significant only for the *LH*-order (see Table 1). To some extent, these results corroborate our hypothesis that the shares of players 1 who do not exert effort and the shares of players 2 who engage in revenge are positively related.

OBSERVATION 4 *On average, revenge is triggered by past untrustworthiness ($\bar{e}$) although the positive correlation between revenge and past $\bar{e}$-shares is at best weakly significant.*

Since we considered a one population model where players could switch roles, the number of monetarily induced players in each matching group varied over periods. It is therefore interesting to investigate whether the choice variables of the monetarily non-induced players hinge on the lagged frequency of subjects with no intrinsic concerns. Taking into account for each treatment the average number of monetarily

induced players in period t, we find that this is unrelated to the average level of no effort or revenge in period $t + 1$ (for both choice variables, $\bar{e}$ and r, in each treatment under either order, the Spearman rank correlation coefficients are positive but not significant; $p > 0.297$ always). Intrinsically motivated participants in our experiment do not seem to be affected by the share of monetarily induced players either at the outset or throughout the session.

Let us now briefly look at the behavior of the monetarily induced subjects, starting with players 1. Recall that we set $n = 9$ for three subjects in the *H*-treatment and $n = 67$ for one subject in the *L*-treatment. Monetarily induced players 1 were present, on average, about half of the times in each treatment under either order. Their exerted average relative frequencies of effort (relative to the numbers of presence, with averages over matching groups and periods) were 0.69 (0.88) and 0.64 (0.95) in the *H*- (*L*-) treatment under the *HL*- and *LH*-order, respectively.

If all players with a positive n were abiding by their monetary incentives, these relative frequencies should be one. But this is not what we observe. In particular, in the *L*-treatment under the *LH*- (*HL*-) order, out of the 6 monetarily induced players – one per population/matching group – 2 (4) complied with their pecuniary incentives throughout the experiment, and the remaining 4 (2) invested effort at least 86% (62%) of the times they assumed player 1's role. As regards the 18 induced players 1 in the *H*-treatment – three per population/matching group – their behavior appears much more heterogeneous: under the *LH*-order, 4 players with $n = 9$ always put in effort, 8 (3) did so at least half (a third) of the times, and the remaining 3 never decided for effort; the respective numbers under the reversed order are 2, 13 (2), and 1.

Statistical comparisons of the same treatment under the two orders reveal that whether the treatment is played in the first or in the last 50 periods has not significant effect on the effort choices of monetarily induced players 1 (both $p > 0.70$; two-sided Wilcoxon rank-sum test). This justifies pooling the data from the same treatment across sessions so as to study the "pure" effects of initial conditions on monetarily induced subjects. As the numbers reported above suggest, the *L*-treatment is significantly more effective in inducing effort of subjects with monetary equivalents

than the *H*-treatment ($p = 0.002$; two-sided Kolmogorov-Smirnov test relying on 12 independent observations). We explain this result by the differences in monetary inducement: while $n - \overline{n}(y) \approx 64$ in the *L*-treatment, we have $n - \overline{n}(y) \approx 9$ in the *H*-treatment. We can, therefore, conclude that monetary inducement does affect behavior, though not always in a perfectly deterministic way. Of course, the weaker inducement of players 1 in the *H*-treatment may also account for the lack of significant difference in non-induced behavior of players 2 between treatments. The evidence on monetarily induced players 1 can be summarized as follows.

OBSERVATION 5 *Regardless of the order of treatments, monetarily induced player 1's decisions depend on monetary incentives for "moral behavior".*

Turning to the retributive behavior of monetarily induced players 2 (three in *H* with $m = 35$, and one in *L* with *m* = 30), the number of times they were actually assigned to this role were, on average, nearly half the feasible ones. The relative average frequencies of revenge were 0.38 (0.36) and 0.26 (0.26) in the *H*- (*L*-) treatment under the *HL*- and *LH*-order, respectively. Like monetarily induced players 1, also players 2 who are predicted to retaliate do not act in accordance with their monetary incentives. However, non-punishing in spite of one's monetary incentives is somewhat more "excusable" than no-effort because punishment may harm an expert who has invested effort. In particular, in the *L*-treatment under the *LH*- (*HL*-) order, except 1 (3) player(s) who sued 81% (at least 57%) of the times she (they) took on this role, the other 5 (3) investors with *m* = 30 sued less than half of the times, with 2 (1) of them never suing. The behavior of the 18 monetarily induced players 2 in the *H*-treatment is no much different: only 2 (1) monetarily induced player(s) abided by pecuniary incentives every time they were (she was) assigned to this role, other 2 (5) sued at least half of the times, and the remaining 14 (12) sued, on average, just 11% (15%) of the times under the *LH*- (*HL*-) order.

There is no significant order effect for players with monetary incentives to punish ($p > 0.589$ in both cases; two-sided Wilcoxon rank-sum tests), whose behavior does not significantly depend on initial conditions ($p = 0.518$ according to a Kolmogorov-Smirnov test comparing *H* and *L* with data pooled across sessions). However, the

frequencies of revenge were significantly larger for induced than for non-induced players 2 in either treatment ($p < 0.02$ in both cases; Kolmogorov-Smirnov test relying on 12 independent observations). The results on monetarily induced players 2 are summarized in our last observation.

OBSERVATION 6 *Regardless of the order of treatments, monetarily induced players 2 react to their incentives, though not to the degree one would expect from deterministic rationality*.

6 Conclusions

We investigated the co-evolutionary stability of trustworthiness and retribution by means of both an indirect evolutionarily model and an experimental study. The latter is evolutionary in that it provided participants with feedback about population behavior: in each period, except the first, participants were told what the proportions of untrustworthy and retributive choices in the previous period were. To examine whether initial conditions affect intrinsic motivations to behave trustworthy and to punish, we induced monetary equivalents of moral concerns and considered two treatments differing in the number of monetarily induced participants (either one or three per player role). Participants with such monetary equivalents should behave "morally" for purely opportunistic reasons.

The main proposition of the indirect evolutionary model, predicting an erosion of intrinsic motivations, is clearly rejected by the experimental analysis: "non-induced" participants remained mostly cooperative, and a positive small percentage of punishment persisted, regardless of whether the share of incentivized moral behavior was large or small, and independently of when a given share was faced. Hence, the experimental evidence garnered here contradicts not only the standard rational choice approach, but also a non-standard methodology modeling the evolution of internal dispositions by means of endogenous preference changes.

Our data also reveal that participants who were monetarily "induced" to behave in certain intended ways did so more often when their population share was lower. Other

individuals' actions mattered in a way which is incompatible with narrowly conceived evolutionary dynamics. Whether the indirect evolutionary approach can provide appropriate tools for modeling such finding may be regarded as an open question – but not entirely so. Assuming that motivations can affect evolutionary success solely in indirect ways is not an inherent constraint of the approach (see, e.g., GÜTH AND PELEG [2001]). Direct success effects of strong regret or strong desire for revenge may be justified by claiming that undeserved gains are used less efficiently or that maintenance of one's self image promotes success. Accordingly, the concept of reproductive success may be modified by considering certain behavioral forms not only more satisfying (i.e., intrinsically rewarding) but also materially effective.

Appendix

Sample instructions (originally in German)

Welcome and thanks for participating in this experiment.

You receive €2.50 for having shown up on time. Please read the instructions – which are identical for all participants – carefully. From now on any communication with other participants is forbidden. If you have any questions or concerns, please raise your hand.

We will answer your questions individually.

The experiment allows you to earn money. How much you will earn depends on your own decisions, on the decisions of others' participants, and on chance. During the experiment, amounts will be denoted by ECU (Experimental Currency Unit). ECU are converted to euros at the following exchange rate: 1 ECU = € 0.1. This means that 100 ECU = € 10.

Detailed instructions

The experiment is divided into several periods. In every period, participants are matched in groups of two. The composition of the groups will change after each period, so that the person you are matched with will be different from one period to the next.

In each two-person group, there is an *expert* and an *investor*. In the following, we shall refer to the expert as member *E* and to the investor as member *I*.

Structure of each period

In each period, member *E* is endowed with 50 ECU, and advises member *I* on an investment, which can be successful or not. The probability of success depends on whether member *E* decides to exert an effort when advising or not. In particular,

- if member *E* exerts effort, the probability that the investment succeeds is 2/3,
- if member *E* does not exert effort, this probability is reduced to 1/3.

Effort costs member *E* 20 ECU.

A successful investment yields another 50 ECU to member *E* and 100 ECU to member *I*.

An unsuccessful investment yields 0 ECU to both group members.

When the investment is unsuccessful, member *I* can sue member *E*.

- If member *I* sues his/her group member, member *I* pays 20 ECU and member *E* loses his/her endowment of 50 ECU.

- If member *I* does not sue his/her group member, member *I* does not pay anything and member *E* keeps his/her endowment.

Member *I* has to prepare for suing before she knows whether the investment actually failed and before she knows whether member *E* exerted effort or not. In other words, when deciding on whether to sue, (s)he does not know whether the investment failed and whether its failure is at least partly due to member E's non-effort.

Thus, each period consists of the following three stages:

1. First, member *E* decides whether or not to exert effort when advising member *I*.
2. Then, member *I* decides whether (s)he will or will not sue member *I* in case of unsuccessful investment.
3. Finally, chance decides whether the investment is successful or not, where the probability of success is 2/3 if member *E* exerted effort, and 1/3 if *E* did not exert effort.

Period-earnings

The earnings in each single period depend on whether the investment is successful or not, member *E* exerts effort or not, and member *I* sues member *E* or not.

► If the investment is **successful**:

● *E*'s earnings in case of effort = 50 (initial endowment) + 50 (earnings from investment) – 20 (effort costs) = *80*

● *E*'s earnings in case of no effort = 50 (initial endowment) + 50 (earnings from investment) = *100*

● *I*'s earnings = *100* (earnings from investment)

► If the investment is **not successful**, and member *E* did not exert effort:

●	*E*'s earnings = *50* (initial endowment) – ***n***	⎫	If member *I* does not sue member *E*
●	*I*'s earnings = *0*	⎭	
●	*E*'s earnings = *0*	⎫	If member *I* does sue member *E*
●	*I*'s earnings = ***m*** – *20* (suing costs)	⎭	

► If the investment is **not successful**, and member *E* did exert effort:

● *E*'s earnings = 50 (initial end.) – 20 (effort costs) = *30* ● *I*'s earnings = *0*			If member *I* <u>does not sue</u> member *E*
● *E*'s earnings = – *20* (effort costs) ● *I*'s earnings = – *20* (suing costs)			If member *I* <u>does sue</u> member *E*

You will learn the value of the two parameters "***m***" and "***n***" at the beginning of the experiment. These values may differ across participants.

The roles of expert *E* and investor *I* are assigned randomly in each period. This means that your role assignment can change from one period to the next. You will read the role you have been assigned to on your computer screen at the beginning of each period.

The information you receive at the end of each period

At the end of each period, you will be informed about

1) your earnings in this period;
2) this period's earnings of the group member you have interacted with;
3) the share of participants (in %) in the role of *E* who have chosen *not to exert effort* in the current period;
4) the share of participants (in %) in the role of *I* who have chosen *to sue* in case of unsuccessful investment in the current period.

Your final earnings

At the end of the experiment, four periods will be randomly selected for payment. Your earnings in these 4 periods will be added up. The resulting sum will be converted to euros and paid out.

Please note that you may make losses. You have to cover them by completing an additional task at the end of the experiment. The additional task requires to search and mark specific symbols in a text. By doing so, you can compensate 1 Euro loss by correctly completing half a page.

REFERENCES

ANDREONI, J., W. HARBAUGH, AND L. VESTERLUND [2003], "The Carrot or the Stick: Rewards, Punishments and Cooperation," *American Economic Review*, 93, 893−902.

AUMANN, R. J. [1981], "Survey of Repeated Games," pp. 11−42 in: R. Aumann, W. Hildenbrand, and J. Rosenmüller (eds.), *Essays in Game Theory and Mathematical Economics*, Bibliographisches Institut: Mannheim.

BRANDTS, J., AND G. CHARNESS [2000], "Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games," *Experimental Economics*, 2, 227−238.

FEHR, E., AND S. GÄCHTER [2000], "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90, 980−994.

FISCHBACHER, U. [1999], "Zurich Toolbox for Readymade Economic Experiments," Working Paper 21, University of Zurich, Switzerland.

GÜTH, S., W. GÜTH, AND H. KLIEMT [2002], "The dynamics of trustworthiness among the few," *Japanese Economic Review*, 53, 369−388.

GÜTH, W., AND H. KLIEMT [2004], "Perfect or Bounded Rationality? Some Facts, Speculations and Proposals," *Analyse & Kritik*, 26, 364−381.

− −, W. LEININGER, AND G. STEPHAN [1991], "On Supergames and Folk Theorems: A Conceptual Analysis," pp. 56−70 in: R. Selten (ed.), *Game Equilibrium Models II. Morals, Methods, and Markets*, Springer: Berlin.

− −, M.V. LEVATI, A. OCKENFELS, AND T. WEILAND [2005], "Buying a Pig in a Poke: An Experimental Study of Unconditional Veto Power," Discussion Paper 39-2005, Max Planck Institute of Economics, Jena, Germany.

− − AND B. PELEG [2001], "When Will Payoff Maximization Survive? - An Indirect Evolutionary Analysis, *Evolutionary Economics*, 11, 479−499.

KREPS, D., P. MILGROM, J. ROBERTS, AND R. WILSON [1982], "Rational Cooperation in the Finitely-Repeated Prisoners' Dilemma," *Journal of Economic Theory*, 27, 245−252.

KREPS, D., AND R. WILSON [1982], "Reputation and Imperfect Information," *Journal*

of Economic Theory, 27, 253−279.

MACKIE, J. L. [1982], "Morality and the Retributive Emotions," *Criminal Justice Ethics*, 1982, 3−10.

MARKOWITZ, H. M. [1952], "Portfolio Selection, " *Journal of Finance,* 7, 77−91.

MASCLET, D., C. NOUSSAIR, S. TUCKER, AND M.-C. VILLEVAL [2003], "Monetary and Nonmonetary Punishment in the Voluntary Contribution Mechanism," *American Economic Review*, 93, 366−380.

SELTEN, R. [1967], "Die Strategiemethode zur Erforschung des Eingeschränkt Rationalen Verhaltens im Rahmen eines Oligopolexperiments," pp. 136−168 in: H. Sauermann (ed.), *Beiträge zur Experimentellen Wirtschaftsforschung*, J.C.B. Mohr: Tübingen.

− − [1998], "Features of Experimentally Observed Bounded Rationality," *European Economic Review*, 42, 413−436.

WESTERMARCK, E. [1906], *The Origin and Development of Moral Ideas*, Vols. I, II, MacMillan: London.

Werner Güth
M. Vittoria Levati
Max Planck Institute of Economics
Strategic Interaction Group
Kahlaische Str. 10
07745 Jena
Germany
E-mail:
gueth@econ.mpg.de
levati@econ.mpg.de

Hartmut Kliemt
Universität Duisburg
Philosophy Department
Campus Duisburg
Lotharstrasse 65
47048 Duisburg
Germany
E-mail:
hartmut.kliemt@t-online.de

Georg von Wangenheim
Universität Kassel
FB 07 Wirtschaftswissenschaften
Nora-Platiel-Strasse 2
34109 Kassel
Germany
E-mail:
g.wangenheim@uni-kassel.de