Werner Güth, Hartmut Kliemt, Stefan Napel

Population-Dependent Costs of Detecting Trustworthiness
- An Indirect Evolutionary Analysis -

Abstract: If the (un)trustworthy are rare, people will talk about them, making their detection more reliable and / or less costly. When, however, both types appear in large numbers, detecting (un)trustworthiness will be considerably more difficult and possibly too costly. Based on Güth and Kliemt (2000) we analyze how the composition of a population of trustworthy, resp. untrustworthy individuals evolves if the cost and reliability of type detection depend on the population composition.

1. Introduction

If virtuous behavior prevails a rare misdeed will draw attention. It will become a matter of gossip and widely known. This knowledge will influence the behavior of others in encounters with the norm violator. If nearly everybody is misbehaving, the rare trustworthy individual may raise a lot of interest, too. This behavior may become widely known as well and trigger responses. In short, bad as well as good conduct may stand out in a crowd of behavior of the other kind. It will easily be observed and thereby causally influence the behavior of observers (see on such mechanisms from a social science point of view Coleman, J. S. (1988), from a normative perspective Urmson, J. O. (1958)).

Since memory capacity is costly and limited there are obvious informational reasons why type detection should be less costly for given reliability or more reliable for given cost: A complete description of the population based on agents' past behaviour (e.g. "all agents except … are trustworthy") requires less memory or is more easily transmitted and accessed if the minority, which needs to be enumerated, is smaller. Identification heuristics of a given complexity (cost), aiming at an incomplete but still useful description of the population by rules of thumb such as "one cannot trust those with yellow scarves (black shoes, etc.)", can ceteris paribus be more accurate the less individuals need to be singled out. Conversely, less complex rules can be used to achieve a desired reliability.

To study the population dependency of detecting virtue we focus on the virtue of being trustworthy, respectively of failing to show this moral quality. By showing trust the trustor aims at reaching a payoff dominant result as compared to the status quo of no trust but makes himself vulnerable to an act of "exploitation" by the trustee. Trustworthiness is modelled as a modification of the preferences of the trustee. As a result of this modification (due to some kind of intrinsic motivation) the trustee evaluates results in ways other than suggested by objective or material outcomes that reflect reproductive success in the context of the evolutionary model. Intrinsic "moral motivations" prevent the trustworthy trustee from exploiting the trustor, whereas the untrustworthy individuals will not refrain from exploitation should they be trusted.

We assume that to limit their risk, trustors can invest in type detection. Utilizing such a technology they receive a more or less reliable signal whose reliability and cost are, however, not constant as in Güth and Kliemt (2000) but rather population dependent.[1] An extension of our analysis explicitly allows the reliability of the signal of another's type to depend on how the population is composed. More specifically, we assume that the signal's reliabilities (one has to distinguish the signal's reliability when resulting from the trustworthy and the untrustworthy) become worse when the relative frequencies of both virtue types converge. Before analyzing this phenomenon we investigate what to expect when constant reliabilities for more symmetrically composed populations require higher costs of detection.

On a more abstract level, our analysis is comparable to evolutionary studies assuming that the rules of the game change when (average) population play changes. So, for instance, Joosten, Brenner and Witt (2003) are studying games whose payoff parameters depend also on past play. In principle, we do the same but do not presuppose such dependency but rather focus on an institutional aspect, the population dependency of reliability and cost of detection, which

---

[1] One could also have assumed that not only the cost of investing in type detection is population dependent (in the sense of being lower the more one type prevails) but that also the strength of preference modifications depends on how the population is composed. If, for instance, feelings of guilt increase when one is the rare untrustworthy, this should stabilize universal trustworthiness. Similarly, if feelings of guilt get weaker when untrustworthiness becomes more widespread, a monomorphic society of potential exploiters should be stable.

the interacting parties could even render unimportant, e.g. in our situation at hand by not investing in type detection.

Section 2 describes the basic setup more formally. The rational decision behavior for all possible compositions of the population with (un)trustworthy individuals is derived in section 3. In the tradition of the indirect evolutionary approach we then assess the (reproductive) success of the (un)trustworthy type. Assuming success monotonic evolutionary dynamics we determine in section 4 the evolutionarily stable population compositions and their basins of attraction. Section 5 concludes.

2. The model

To capture the trust problem in social interaction, we rely on the trust game in Figure 1 with the same parameter normalization as in Güth and Kliemt (2000) to render our results comparable, where $1 > r > s > 0$.

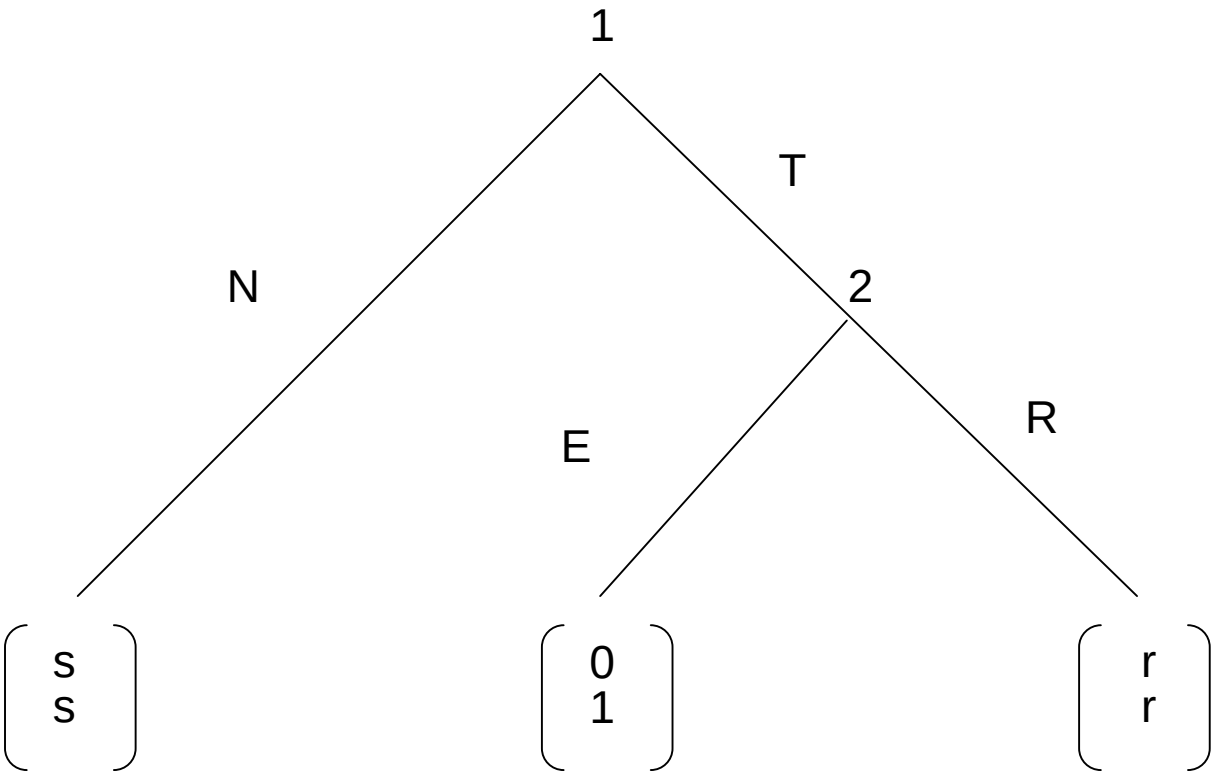

Figure 1

The interpretations of the moves are

N – no trust (in player 2),

T – trust (in player 2),

E – exploiting (player 1's trust),

R – rewarding (player 1's trust).

The payoffs in Figure 1 (top player 1, bottom player 2) represent material or reproductive success. It is assumed that individuals can evaluate results (plays of the game) not only in "objective" terms. In the second–mover role their behavioral choices can be guided by preferences other than furthering reproductive success. It is the presence or absence of such preferences of sufficient strength that renders an individual trustworthy or untrustworthy, respectively. Trustworthiness is captured by a purely intrinsic payoff component m (see Figure 2) with $m = \underline{m} < r - 1$. Correspondingly, we assume for the untrustworthy type of player 2 $m = \overline{m} > 1 - r$. What will be analyzed in the following is the evolution of the

population share *p* of types $m = \underline{m}$, i.e. of those whom one wants to trust in the trust game of Figures 1 and 2.

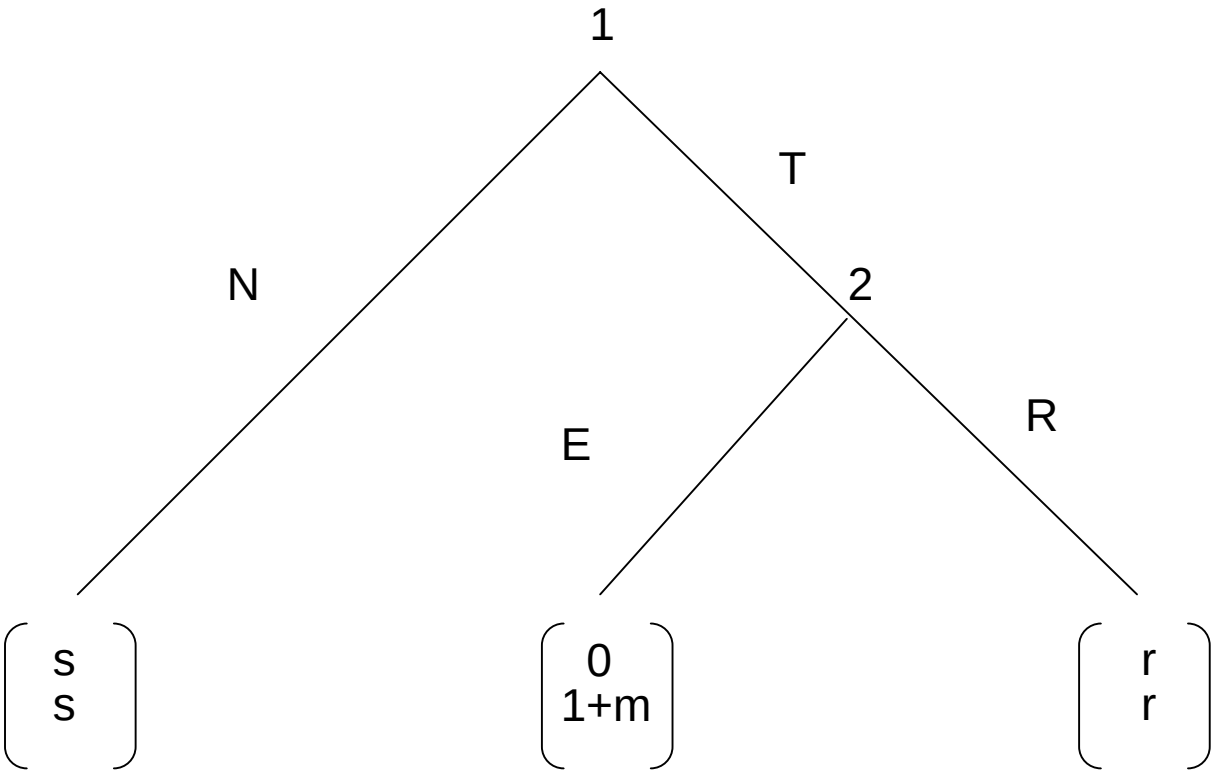

Figure 2

The actual play of the trust game is embedded in a more complex decision process. Assuming an infinite population with random matching (for an alternative see Güth, Güth and Kliemt, 2002) the selected pair of individuals *i* and $j(\neq i)$ confronts the following decision process:

- The two individuals independently decide between investing (*y*), resp. not investing (*n*) in type detection where cost

$$C(p) = \frac{1}{2}\left[kp(1-p) - \frac{1}{6}k + c \right] \text{ with } 6c \geq k > 0,$$

of choosing *y* is population dependent. Clearly, such a cost function has the property that rare types are more cheaply found out due to $p(1-p) \to 0$ both for $p \to 0$ as well as for $p \to 1$. Parameter *k* scales sensitivity of the positive detection cost to the population composition (higher *k* corresponds to greater population dependence), while keeping average cost constant to $\int_0^1 C(p)\,dp = \frac{c}{2}$.

- Chance assigns roles independently of player type, i.e. either individual *i* becomes player 1 and *j* player 2 or vice versa, each with probability 1/2.

- Player 1 decides between *N* (no trust, which would end the interaction with what may be seen as the status quo payoffs) or *T* (trust, which may be seen as an invitation to co-operate). If player 1 has chosen *y* before, he can base his decision on a type signal $\hat{m}$ of player 2's true *m*-type. The reliability of that signal is determined by two parameters $\text{Prob}\left(\hat{m}=\overline{m}\,|\,m=\overline{m}\right)=\overline{\mu}\in\left(1/2,1\right]$ and $\text{Prob}\left(\hat{m}=\underline{m}\,|\,m=\underline{m}\right)=\underline{\mu}\in\left(1/2,1\right]$ meaning that a truly untrustworthy $m=\overline{m}-\text{type}$ is revealed by $\hat{m}=\overline{m}$ with probability $\overline{\mu}$ larger than 1/2 whereas the misleading signal $\hat{m}=\underline{m}$ results with the complementary probability. Similarly, for a true type $m=\underline{m}$ the signal $\hat{m}=\underline{m}$ is more likely than the misleading one.

- In case of 1's decision for *T*, player 2 finally chooses between *E* and *R*.

Note that modifying the order of moves and letting players decide between *y* and *n* when actually being in the role 1 of trustor would merely divide detection costs *C*(*p*) by 2, without any other changes. In the setting envisioned here it may seem more natural, though, to assume that detection costs are borne as a kind of sunk cost beforehand. People either bear the costs of following up what is going on in the group or not. When they by chance encounter a potential partner they must decide "on the spot" whether or not to engage him in a co-operative venture by showing trust or not.

The payoffs are the ones in Figure 2 minus the costs *C(p)* of type detection for the individual(s) having chosen *y*. These (phenotypical) payoffs determine the optimal decision behavior in the process above which will be derived in section 3. Compared to this the composition of types, i.e. the evolution of the population composition parameter $p\in\left[0,1\right]$, is governed by the (genotypical) success of the $\overline{m}$, resp. $\underline{m}-\text{types}$ as determined by rational interaction. This success follows from payoff by setting *m = 0*. We will analyze the evolutionarily stable population compositions *p* in section 4.

3. Rational play as depending on the population composition

Player 2's behavior will depend on his type whenever he is asked to move, i.e. after the move *T* by player 1. More specifically, an $\overline{m}$-type would choose *E* and an $\underline{m}-\text{type}$ *R*.

After n, i.e. when not having invested in type detection, player 1 therefore chooses *T* (yielding *pr*) rather than *N* (yielding *s* for sure) when $p \geq s/r$, and *N* otherwise.

After y, i.e. when having received a signal $\hat{m}$ about player 2's m-type, player 1 will follow the recommendation of a signal $\hat{m} = \underline{m}$ and choose *T* provided that

$$p \geq \frac{(1-\overline{\mu})s}{\underline{\mu}(r-s)+(1-\overline{\mu})s} := RHS$$

where the right-hand side of the inequality is smaller than *s/r* due to $\overline{\mu}, \underline{\mu} > 1/2$ and $s<r$.

If *p* is below this threshold level, even a 'good' signal $\hat{m} = \underline{m}$ pointing towards the trustworthiness of player 2 cannot convince player 1 given her pessimistic initial beliefs about the chances that trust will be rewarded. Similarly, for a sufficiently optimistic prior (large *p*), even a 'bad' signal $\hat{m} = \overline{m}$ cannot dissuade player 1 from trusting. Given her updated beliefs after $\hat{m} = \overline{m}$ she chooses *N* only provided that

$$LHS := \frac{\overline{\mu}s}{\overline{\mu}s+(1-\underline{\mu})(r-s)} \geq p.$$

where the left-hand side above is larger than *s/r* due to to $\overline{\mu}, \underline{\mu} > 1/2$ and $s<r$.

Costly detection activity can be profitable only if the signal is not discarded, i.e. if its recommendation is followed always – and not only when it matches the intended action based on the prior. Thus further analysis of investment will focus on intermediate values of *p* satisfying both of the above conditions (see Güth and Kliemt, 2000, Lemma 3.1). Such values exist since the RHS and LHS are smaller and larger than *s*/*r*, respectively.

Now consider the initial choice between *y* and *n*. Optimal behavior *after n* yields the payoff expectation

$$\left.\begin{array}{ll} pr & p \geq s/r \\ & \text{for} \\ s & \text{for } p < s/r \end{array}\right\}$$

conditional on being assigned to the role of player 1.

Choosing *y* and afterwards always following the recommendation yields

$$\frac{1}{2}p\underline{\mu}r+\frac{1}{2}\Big[p\left(1-\underline{\mu}\right)+\overline{\mu}\left(1-p\right)\Big]s-C\left(p\right) \qquad (*)$$

plus a constant term capturing payoff in case the considered agent is allocated to the role of player 2 (which cannot be influenced by the agent's *n* or *y*-decision). Investigating when (*) exceeds *pr*/2 for p>s/r, and *s*/2 respectively for p<s/r, it follows that *y* is better than *n* (or at least as good) for the subinterval

$$\overline{p}\left(C\left(p\right)\right)\geq p\geq \underline{p}\left(C\left(p\right)\right)$$

of $LHS \geq p \geq RHS$ with

$$\overline{p}\left(C\left(p\right)\right)=\frac{\overline{\mu}s-2C\left(p\right)}{\overline{\mu}s+\left(1-\underline{\mu}\right)\left(r-s\right)}$$

(derived from case $pr \geq s$) and

$$\underline{p}\left(C\left(p\right)\right)=\frac{\left(1-\overline{\mu}\right)s+2C\left(p\right)}{\underline{\mu}\left(r-s\right)+\left(1-\overline{\mu}\right)s}$$

(derived from case $pr < s$). It is possible that $\overline{p}\left(C\left(p\right)\right)<\underline{p}\left(C\left(p\right)\right)$, and then investing in type detection (the decision *y*) is necessarily suboptimal. This case arises whenever

$$2C\left(p\right)>\left(\underline{\mu}+\overline{\mu}-1\right)\left(r-s\right)\frac{s}{r}.$$

However, if average cost *c*/2 and population sensitivity *k* are not too large, $\overline{p}\left(C\left(p\right)\right)\geq p\geq \underline{p}\left(C\left(p\right)\right)$ will be satisfied for an entire interval of population compositions.[2] This is illustrated in Figure 3. The triangle depicts $\overline{p}\left(\kappa\right)$ and $\underline{p}\left(\kappa\right)$ for different detection cost levels κ (in the range $\overline{p}\left(\kappa\right)\geq \underline{p}\left(\kappa\right)$, given *r*=0.8, *s*=0.4, $\underline{\mu}=\overline{\mu}=0.85$), and different

---

[2] A necessary and sufficient condition for this is that $\overline{p}\left(C\left(p\right)\right)>\underline{p}\left(C\left(p\right)\right)$ holds at *p*=s/*r*, i.e. at the peak of the $\underline{p}\left(\kappa\right)$-$\overline{p}\left(\kappa\right)$-triangle in Figure 3. This amounts to $k\frac{s}{r}\left(\frac{r-s}{r}\right)-\frac{k}{6}+c<\left(\underline{\mu}+\overline{\mu}-1\right)\left(r-s\right)\frac{s}{r}$, where the left-hand side can be made arbitrarily small through an appropriate choice of *c* and *k*.

curves $\kappa = C(p)$ illustrate how increasing population sensitivity ($k \in \{c, 6c\}$ with c=0.1) affects the range of p such that indeed $\overline{p}(C(p)) \geq p \geq \underline{p}(C(p))$.

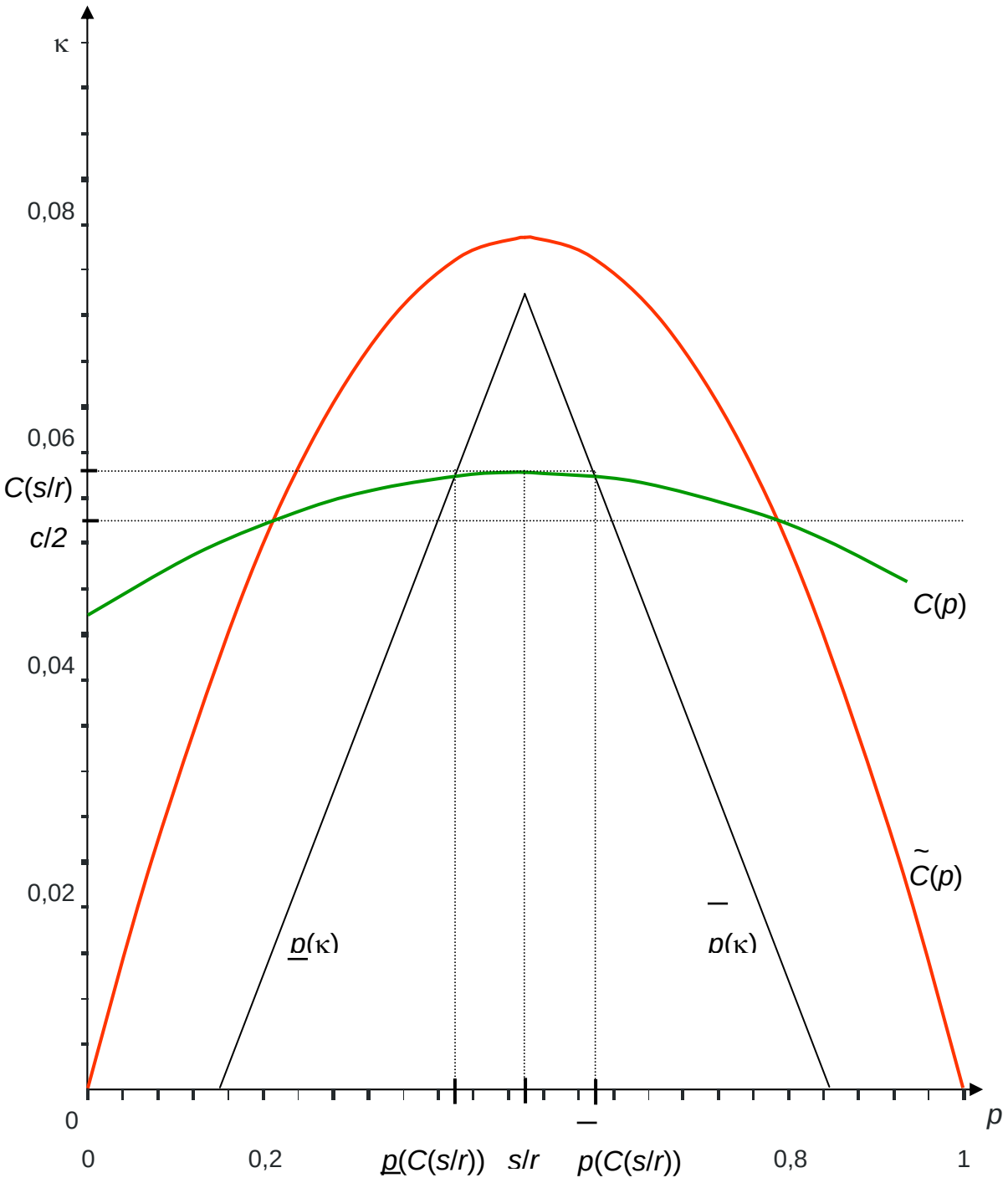

Figure 3

4. The evolution of the population composition

Whenever p violates $\overline{p}\left(C\left(p\right)\right) \geq p \geq \underline{p}\left(C\left(p\right)\right)$, i.e. not investing in type detection (n) is optimal, T will be chosen by player 1 if $p \geq s/r$ and N otherwise. Thus for $p \geq s/r$ an $\underline{m}-$type will earn r materially whereas an $\overline{m}-$type earns more and is thus more successful. Hence for $p \geq s/r$, any monotonic evolutionary dynamics imply that p decreases as long as $p \geq s/r$ (and n is optimal).

Suppose that due to this decrease, p, at some point, starts to satisfy $p < s/r$. Then player 1 chooses N, and both m-types fare equally. But even then, if there are "trembles" in the sense of rare unintentional choices by player 1, the decline of p will continue (see Selten, 1983, 1988, for the justification of such rare trembles). This goes on until either $p^* = 0$ is reached or until p arrives in the range where $\overline{p}\left(C\left(p\right)\right) \geq p \geq \underline{p}\left(C\left(p\right)\right)$ and y becomes optimal.

So, consider population compositions p satisfying $\overline{p}\left(C\left(p\right)\right) \geq p \geq \underline{p}\left(C\left(p\right)\right)$, i.e. player 1 invests in type detection and follows the signal $\hat{m}$ which he receives. Here the material payoff of player 2 and reproductive success of the individual assigned to this role depends on his m-type as follows:

$$\underline{\mu}r + \left(1-\underline{\mu}\right)s - C\left(p\right) \text{ for } m = \underline{m}$$

$$1 - \overline{\mu} + \overline{\mu}s - C\left(p\right) \text{ for } m = \overline{m}.$$

Thus the trustworthy $\underline{m}-$type fares better than the unreliable $\overline{m}-$type of player 2 if

$$\frac{\underline{\mu}}{1-\overline{\mu}} > \frac{1-s}{r-s}, \qquad (**)$$

and vice versa if the opposite inequality applies (we abstract in the following from the degenerate case of equality). Both possibilities are satisfied in generic parameter regions. So, for instance, inequality (**) typically is (not) true for $\overline{\mu}$ close to 1 $\left(\underline{\mu}\right.$ and $\overline{\mu}$ close to $\left.1/2\right)$.

The reverse of (**) is true when $\overline{m}-$types are likely to be mistaken for a trustworthy $\underline{m}-$type (low $\overline{\mu}$) and can then realize a substantial gain (high 1-s). Then $\overline{m}-$types fare universally better and p will sooner or later – faster when $\overline{p}\left(C\left(p\right)\right) \geq p \geq \underline{p}\left(C\left(p\right)\right)$ or $p>s/r$, and slower otherwise – decrease till reaching $p^* = 0$ which is for these parameter configurations the only evolutionarily stable population composition.

If, however, inequality (**) holds, then p increases in the range $\bar{p}\left(C(p)\right) \geq p \geq \underline{p}\left(C(p)\right)$, which depends on p due to the population dependency of $C(p)$, and decreases outside this range. This, of course, means that for all initial population compositions $p_0 < \underline{p}\left(C(p_0)\right)$ or $p_0 > \bar{p}\left(C(p_0)\right)$ one starts out with a decrease of p over time whereas for $\bar{p}\left(C(p_0)\right) \geq p_0 \geq \underline{p}\left(C(p_0)\right)$ one starts with an increase of p over time. In the latter case this process will finally lead to a stable population composition p satisfying

$$\bar{p}\left(C(p)\right) = p$$

or

$$kp^2 - \left[\bar{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k\right]p + \bar{\mu}s - c + \frac{k}{6} = 0.$$

Note that $\bar{p}\left(C(p)\right)$ decreases (increases) for $p<1/2$ ($>1/2$), and $\bar{p}\left(C(0)\right) = \bar{p}\left(C(1)\right) < 1$. Therefore, of the two solutions of the quadratic equation

$$p = \frac{\bar{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k}{2k} \pm \frac{\sqrt{\left[\bar{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k\right]^2 - 4k\left(\bar{\mu}s - c + \frac{k}{6}\right)}}{2k},$$

only the smaller one qualifies as an evolutionarily stable population composition.

How does one actually determine the direction in which p changes (up or down) when considering a time point t=0where the population composition is $p_0 \in [0,1]$? Let us describe this for the more interesting case (**) where p would increase in the interval $\left(\underline{p}\left(C(p_0)\right), \bar{p}\left(C(p_0)\right)\right)$. The first thing to check is whether $2C(p_0) > \left(\underline{\mu} + \bar{\mu} - 1\right)(r-s)\frac{s}{r}$ holds at all. If not, the interval $\left(\underline{p}\left(C(p_0)\right), \bar{p}\left(C(p_0)\right)\right)$ is empty and p would decrease (fast or slow) throughout. If the condition holds, however, we know that $p_0 \in \left(\underline{p}\left(C(p_0)\right), \bar{p}\left(C(p_0)\right)\right)$ and that p will increase from p_0 to a new level p_t, for which one repeats the analysis.

So, if $\bar{p}\left(C(p_0)\right) \geq p_0 \geq \underline{p}\left(C(p_0)\right)$, the population composition converges from below to

$$\bar{p}^* = \frac{\bar{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k - \sqrt{\left[\bar{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k\right]^2 - 4k\left(\bar{\mu}s - c + \frac{k}{6}\right)}}{2k}.$$

If instead

$$1 \geq p_0 > \overline{p}\left(C\left(p_0\right)\right),$$

then p converges to $\overline{p}^*$ from above because for $p > \overline{p}\left(C\left(p_0\right)\right)$ nobody invests in type detection (i.e. chooses *n*). Now $p_0 > \overline{p}\left(C\left(p_0\right)\right)$ is equivalent to

$$kp_0^{\,2} - \left[\overline{\mu}s + \left(1-\underline{\mu}\right)\left(r-s\right) + k\right]p_0 + \overline{\mu}s - c + \frac{k}{6} < 0$$

or $p_0 > \overline{p}^*$. In view of this, the basin of attraction for $\overline{p}^*$ is $p_0 > \underline{p}\left(C\left(p_0\right)\right)$, whereas the basin of attraction for $p^* = 0$ is $p_0 < \underline{p}\left(C\left(p_0\right)\right)$.

Condition $p_0 > \underline{p}\left(C\left(p_0\right)\right)$ (noting that $\underline{p}\left(C\left(p\right)\right)$ first increases and then decreases in p and that $\underline{p}\left(C\left(0\right)\right) = \underline{p}\left(C\left(1\right)\right) > 0$) can also be expressed as

$$p_0 > D := \frac{k - \underline{\mu}\left(r-s\right) - \left(1-\overline{\mu}\right)s + \sqrt{\left[k - \underline{\mu}\left(r-s\right) - \left(1-\overline{\mu}\right)s\right]^2 + 4k\left[s\left(1-\overline{\mu}\right) + c - \frac{k}{6}\right]}}{2k}.$$

Accordingly there exists a threshold D determining whether an initial population composition p_0 leads to an $\overline{m}-$monomorphism or $p^* = 0$, namely for $p_0 < D$, or to a bimorphic population composed of a $\overline{p}^*$-share of $\underline{m}-$types and a complementary $1-\overline{p}^*-$share of $\overline{m}-$types, namely when $p_0 > D$.

Dynamics for given population-dependent costs $C(p)$ are illustrated in Figure 4. The solid line indicates comparatively "fast" movement, corresponding to a strict payoff (dis)advantage of trustworthy agents. Movement along the dotted line is "slow" because it is driven by mutations, i.e. agents in the role of player 1 who by mistake trust and then make trustworthy agents in the role of player 2 fare worse than others (who take advantage of the mistake).

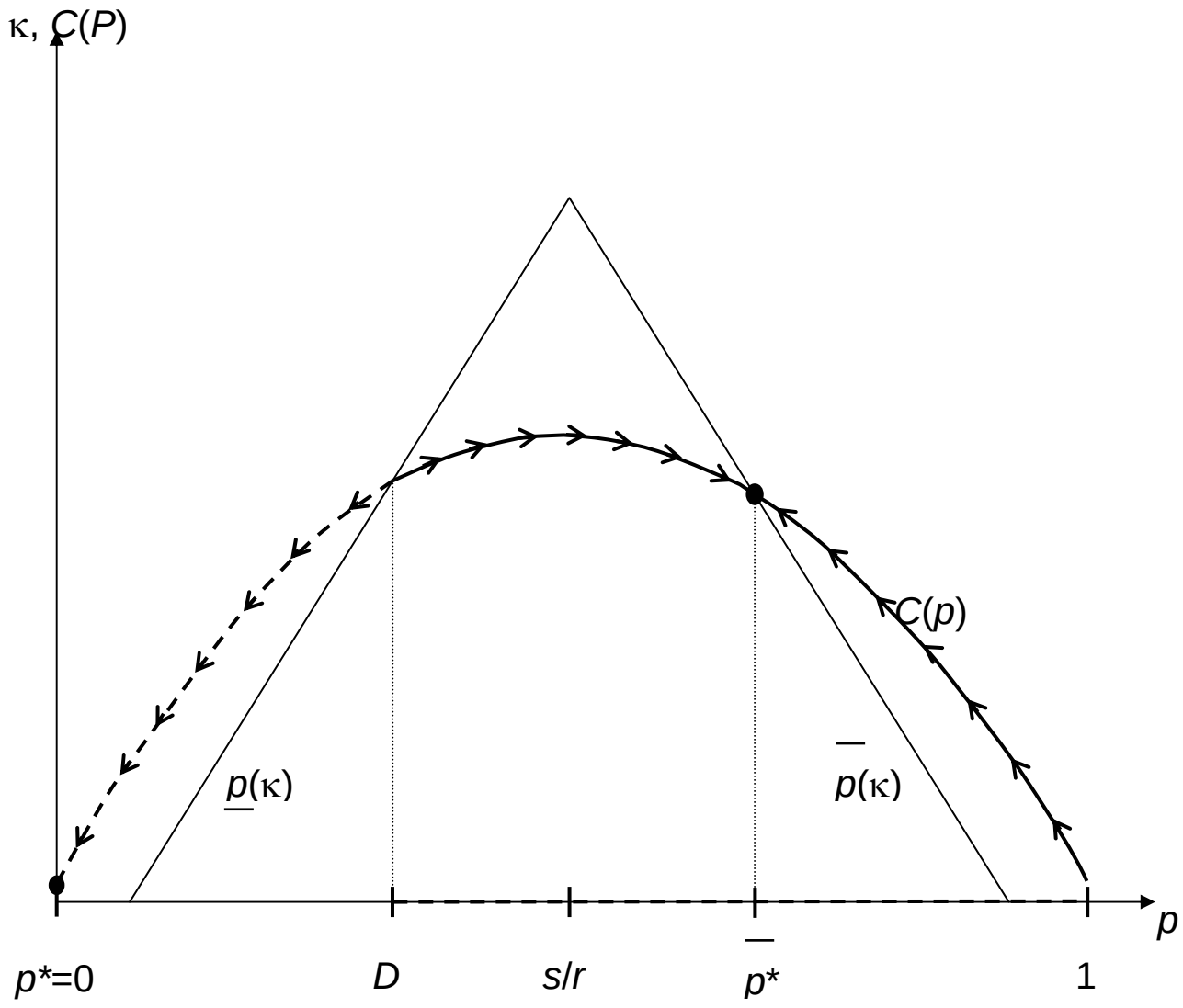

Figure 4

It can easily be seen that a stronger sensitivity of investment costs $C(p)$ on the population composition, i.e. a higher coefficient k, increases D and thus the basin of attraction of $p^* = 0$. At the same time, it also decreases $\overline{p}^*$. Thus, if the costs of type detection rise faster as the rarer m-type gets less rare (cost are "more" population dependent), the chances of a bimorphic population are worsened and the bimorphic population will on average be less

virtuous. The effect of c, i.e. of the fixed cost parameter, has already been discussed by Güth and Kliemt (2000, see also their discussion of the reliability parameters $\underline{\mu}$ and $\overline{\mu}$).

5. Extensions

Above baseline model of population-dependent detection costs lends itself to a number of variations and extensions. First, it was based on a specific, simple functional form of $C(p)$. Many plausible alternatives to the quadratic shape (which has great analytical convenience) exist. One example is a bell shaped form that reflects that costs may initially increase only slowly as more and more (un)trustworthy individuals are added to a population dominanted by $\underline{m}-$types ($\overline{m}-$types). This case in fact allows for multiple stable polymorphic population states, as illustrated in Figure 5, where of the six intersection points of $C(p)$ and the triangle only $\overline{p}_1^*$, $\overline{p}_2^*$ and $\overline{p}_3^*$ are stable bimorphisms with generic basins of attraction (indicated by the direction arrows on $C(p)$), whereas $\underline{p}_1$, $\underline{p}_2$ and $\underline{p}_3$ are just the watersheds separating the basins of attraction.

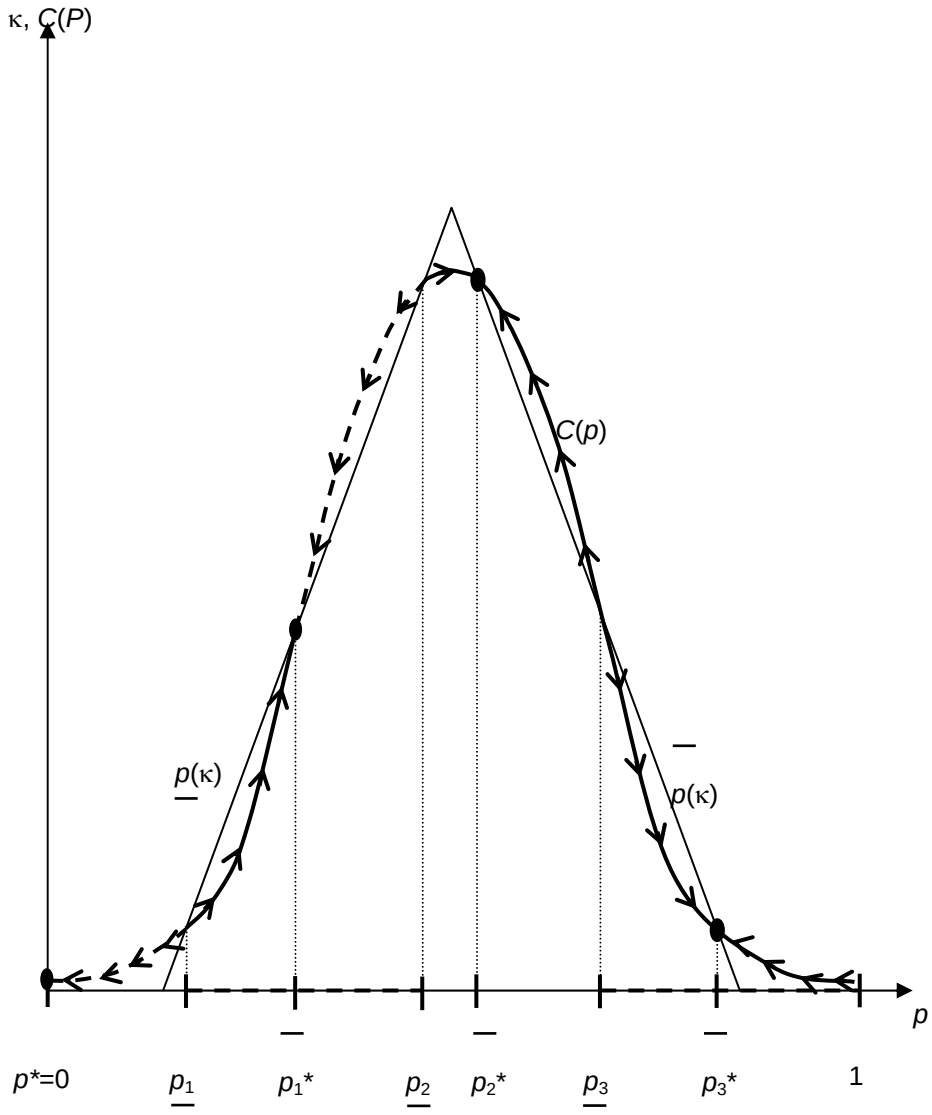

Figure 5

Second, as already indicated in the Introduction, population-dependent costs of detection can be viewed as capturing in an indirect way the population dependency of *signal reliabilities*. So an alternative setup of the model would have taken detection costs to be fixed at some level c but assumed that reliability parameters $\underline{\mu}$ and $\overline{\mu}$ decrease (in possibly asymmetric fashion) as the population state p approaches its least informative level of ½. Different plausible versions of population-dependent reliabilities $\underline{\mu}(p)$ and $\overline{\mu}(p)$ could then be

considered. While many of them would simply induce a rounded version of the $\overline{p}(\kappa)$ and $\underline{p}(\kappa)$-triangle in Figures 3–5 (with evolution of p along a horizontal line $C(p)=\kappa$), new phenomena may arise. In particular, it is possible that there are more than two preference reversals regarding the decision of whether to invest in type detection (y) or not (n) as the share of trustworthy agenst, p, increases from 0 to 1. This is illustrated in Figure 6 (using parameters r=0.7, s=0.35, k=1.8, and μ^{max}=0.99) for the case of

$$\underline{\mu}(p)=\overline{\mu}(p)=\begin{cases} max\left\{\mu^{max}-kp^2,\frac{1}{2}\right\} & \text{if } p\le\frac{1}{2},\\ max\left\{\mu^{max}-k(1-p)^2,\frac{1}{2}\right\} & \text{if } p>\frac{1}{2}. \end{cases}$$

These (symmetric) reliabilities fall from a maximal level of μ^{max} at p=0 and p=1 to μ^{max} - k/4 or ½, whichever is larger, as p=1/2 is approached. So the probability $1-\underline{\mu}$ of falsely expecting an untrustworthy $\overline{m}-$type in case of an $\underline{m}-$encounter is lower (higher) when $\overline{m}-$types are rare (close to ½-population share), and the probability $1-\overline{\mu}$ of falsely expecting an untrustworthy $\underline{m}-$type in case of an $\overline{m}-$encounter is smaller (larger) when $\underline{m}-$types are rare (close to ½-population share).

For the simple cost of $C(p)\equiv\overline{C}$ with $\overline{C}$ small enough to lie below the two peaks (see Figure 6 – note that, in general, the two peaks can have different heights) the result can be described as follows:[3] the two stable bimorphisms $\overline{p}_1^*$ and $\overline{p}_2^*$ have generic basins ($\underline{p}_1$, $\underline{p}_2$) and ($\underline{p}_2$, 1], respectively; i.e., $\underline{p}_1$ and $\underline{p}_2$ are again watersheds separating the basins of attraction of $\overline{p}_1^*$, $\overline{p}_2^*$ and the stable monomorphism p^*=0 with [0, $\underline{p}_1$) as its basin of attraction.

One can obviously complicate the analysis and create more stable bimorphisms, e.g., by combining a bell-shaped cost function as in Figure 5 with two-peaked "triangles" as in Figure 6. In our view, the more important conclusion is that by allowing population-

[3] We do not explicitly distinguish between fast and slow decline of p above the camel-shaped curve indicating gross benefits from exploiting type signals: the decline is driven by mutations only for $s/r<p$, while for $s/r>p$ there is a strict disadvantage for $\overline{m}-$types even without rare mutations.

dependent type detection costs and / or reliabilities of type signals,[4] the phenomenon of multiple stable bimorphisms can arise. It is thus possible that equally structured societies (e.g. those with more or less equal payoff parameters as defining the interaction structure in Section 2) reveal different positive population shares of (un)trustworthy individuals, solely since they started out differently. It also suggests a new kind of policy for improving the trustworthiness of a society which, in view of our analysis, we can describe as "watershed jumping". A policy measure should aim at restarting the evolutionary *p*-process above the lower watershed of the better bimorphism. Note, however, that we cannot justify a p=1-monomorphism (except by trivially assuming $\overline{\mu} = \underline{\mu} = 1$ and $C(1)=0$), see also Fn. 1 above), and can never rule out a p^*=0-monomorphism. It seems that we cannot live without some untrustworthy ones and have to expect only them when starting out with too few trustworthy members of society. If so there is, however, the chance of "watershed jumping", i.e. of policy trying to inspire an evolutionary increase of *p*, which might have to be repeated to reach the best possible bimorphism in the sense of a maximally stable population share of trustworthy individuals.

---

[4] Another alternative (with similar qualitative implications as the cases already considered) would be to allow agents to optimally *choose* the desired individual signal reliabilities according to a cost function $C\left(\overline{\mu},\underline{\mu},p\right)$ which is increasing in $\overline{\mu}$ and $\underline{\mu}$, and decreasing in distance $\left|\frac{1}{2} - p\right|$.

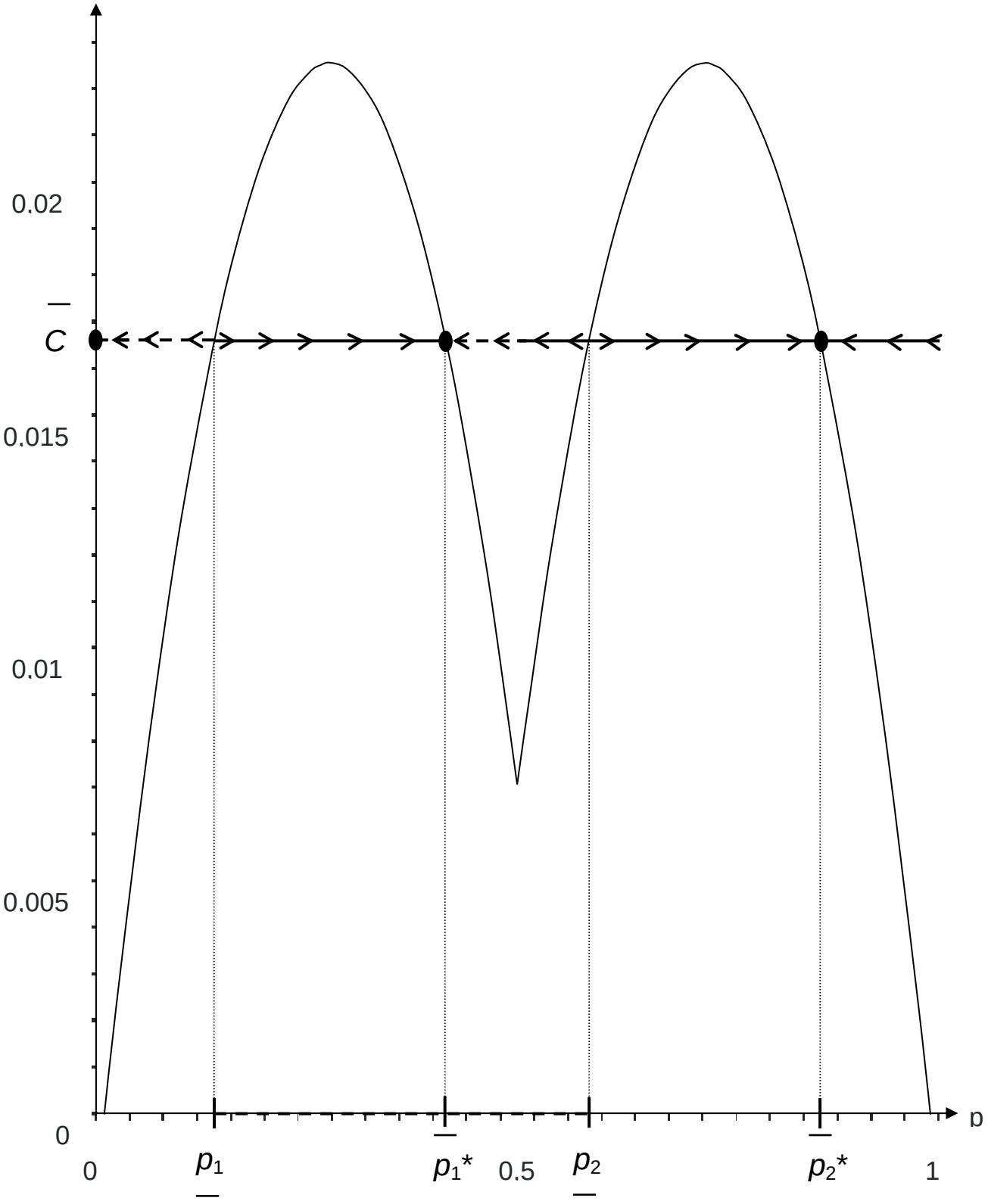

Figure 6

6. Conclusions

Quite generally, the resources of a habitat will depend on how it is inhabited (possibly by more than just one species (see Aumann and Güth, 2000). If, for instance, a habitat is overused, it may not be sustainable what might even endanger the species which rely on it.

Here the habitat changes concern only how costly it is to obtain a signal on the other's (un)trustworthiness and how reliable the signals are. Our assumption that type detection becomes easier the more monomorphic the population could be turned around by stating that the habitat becomes more difficult when both, the trustworthy and the untrustworthy ones, are equally numerous. In this sense our rather specific analysis can be more broadly interpreted as one where the population composition and the quality of the habitat are coevolving. Unlike in the former study of Güth and Kliemt (2000) this allows for multiple stable bimorphisms and enriches the spectrum of explanations why societies may differ in the type composition and of possible policy measures to improve the trustworthiness in society.

7. References

Aumann, Robert and Güth, Werner. 2000. Species survival and evolutionary stability in sustainable habitats - The concept of ecological stability. Journal of Evolutionary Economics, 10, 437-447

Coleman, J. S. (1988): "Free Riders and Zealots: The Role of Social Networks," Sociological Theory, 6, 52-57.

Güth, Sandra, Güth, Werner and Kliemt, Hartmut. 2002. The Dynamics of Trustworthiness Among the Few. The Japanese Economic Review, vol. 53/4, 369-388.

Güth, Werner and Kliemt, Hartmut. 2000. Evolutionary stable co-operative commitments. Theory and Decision, 49, 197-221.

Joosten, Reinoud, Brenner, Thomas and Witt, Ulrich. 2003. "Games with frequency-dependent stage payoffs", International Journal of Game Theory, 31, 609-620.

Selten, Reinhard (1983): "Evolutionary stability in extensive two-person games", Mathematical Social Science 5: 269-363.

Selten, Reinhard (1988): "Evolutionary stability in extensive two-person games – correction and further developments", Mathematical Social Science 16: 223-266.

Urmson, J. O. (1958): "Saints and Heroes," in Essays in Moral Philosophy., ed. by I. Melden. Seattle/London: University of Washington Press, 198 ff.