

final version:

Applied Artificial Intelligence, Volume 14, Number 8, pp 769 - 784, Taylor and Francis, 2000

# Boosting Cooperation by Evolving Trust

Andreas Birk

Vrije Universiteit Brussel, Artificial Intelligence Laboratory,

Pleinlaan 2, 10G725, 1050 Brussels, Belgium

birk@ieee.org

## Abstract

Instead of establishing trust through defining compliance-based standards like protocols augmented by cryptographic methods, it is shown that trust can emerge as a self-organizing phenomenon in a complex dynamical system. It is assumed that trust can be modeled on basis of an intrinsic property called trustworthiness in every individual  $i$ . Trustworthiness is an objective measure for other individuals whether it is desirable to engage in an interaction with  $i$  or not. Trustworthiness can not directly be perceived. Building trust therefore relates to estimating trustworthiness. Subjective criteria like outer appearance are important for building trust as they allow to handle unknown agents for whom data from previous interactions does not exist. Here, trustworthiness is grounded in the strategies of agents who engage in an extended version of the iterated Prisoner's Dilemma. Trust is represented as preference to be grouped together with agents with a certain label to play a game. It is shown that stable relations of trust can emerge and that the co-evolution of trust boosts the evolution of cooperation.

## 1 Introduction

The investigation of the formation and the application of trust is interesting from two different perspectives. First, it relates to basic research on fundamental principles of social interactions between living systems, especially humans. Second, constructive approaches investigating trust are important for applications, allowing autonomous interactions between artificial systems.

Almost all higher life-forms interact with other individuals of their kind, leading to complex social behaviors. Man is no exception, in contrary, social and cultural behaviors are among the most crucial aspects of the major pride of humans, namely cognition. The study of the social interactions of living systems in general and of humans in particular is accordingly an important issue of basic research, and it is pursued in many different fields like ecology, economics and social science, to name just a few.

Today, there is an increasing amount of artificial autonomous systems, which control various devices without a continuous or explicit supervision by humans. To reach their full potential, artificial autonomous systems must be capable of interacting with each other. E-commerce can for example only be successful if a multitude of devices cooperates in an autonomous delivery of goods *into*

the buyer's place. There must be for example a goods-port, which is capable of handling deliveries without the presence of a human supervisor.

The classic scientific fields investigating the basic principles of social interactions between living systems were forced to rely on *descriptive* approaches for their work, as the manipulation of living societies is simply infeasible or even immoral. The appearance of artificial autonomous systems now leads to the possibility as well as to the need to use *constructive* approaches. It is on the one hand a gift, as it allows to "tinker" with artificial but nonetheless complex societies to investigate basic research questions. It poses on the other hand a serious challenge as further technological progress needs working solutions for concrete applications.

The field of Multi-Agent-Systems or short MAS [GB99, JW98a, JW98b, CW94, DM91] focuses on compliance-based approaches for a constructive investigation or exploitation of artificial societies. This means, this field tries to establish standards for agent languages and architectures [MSR99, SRW98, MJW96, WMT96] within which interactions take place. In respect to trust, cryptographic methods are for example used [LN98, Har98, Phi97], which establish a well-defined security at the cost of restricting interactions to systems which comply to the standard.

Here in contrast, trust is formed in a dynamical process. There is no absolute security as trusted systems can cheat. But the process is completely open and robust as trust is not predefined, but it emerges from subtle interactions between the systems. The basic ideas of this process go back to two roots, namely the field of Artificial Life or short Alife [Ste94a, LTFR90, Lan89] and the field of Evolutionary Game-Theory [Smi84, Axe84, AH81, SP73]. Before the process of the formation of trust can be described, it is necessary to first define the notion of trust itself as it is used here.

The basis for trust is an intrinsic property of each individual in form of the so-called **trustworthiness**. The trustworthiness of an agent  $a_A$  is an *objective* measure for an other agent  $a_B$  of the desirability of interactions with  $a_A$ . If the, possibly continuous, trustworthiness of  $a_A$  is high, it is highly desirable for  $a_B$  to engage in trust-based interactions with  $a_A$ . Vice versa, if the trustworthiness of  $a_A$  is low, it is highly undesirable for  $a_B$  to engage in interactions with  $a_A$ .

The trustworthiness of  $a_A$  can be dynamic, both in time as well as in respect to the agent-space, i.e., the trustworthiness of  $a_A$  for agent  $a_B$  can be different to the trustworthiness of  $a_A$  for agent  $a_C$  at the same moment in time. The problem of trustworthiness is that it is an internal state of agent  $a_A$ , which can not be accessed by any other agent. It might even be based on hidden or so-to-say unconscious processes such that  $a_A$  himself can not truly determine its own trustworthiness for others. In addition, trustworthiness is dynamic, i.e., even when correctly determining it once in respect to its meaning for one agent, this information can be useless shortly after in time or in respect to an other agent.

Any process which tries to establish an approximation of the trustworthiness of  $a_A$  is here denoted as building **trust**. Processes for building trust often include a non-rational component in the sense that decisions on how to deal with an other individual are not only based on previous interactions with this individual, but also on other, presumably *subjective* criteria. These criteria for example include outer appearance, recommendations from others and so on. Subjective processes for building trust are extremely important as they allow decisions whether to interact or not with unknown individuals, i.e., individuals who have not been encountered in previous interactions.

There are more or less unlimited possibilities for the representation of trustworthiness and for the implementation of trust building processes. In a purely descriptive approach to the matter, Bacharach and Gambetta for example propose in [BG00] to use certain properties of pay-off matrices for representing trustworthiness and signaling theory to formalize the aspects of a subjective

building of trust. In this article, trustworthiness is grounded itself in the strategies of agents who engage in an extended version of the iterated Prisoner’s Dilemma. Trust is represented as preference to be grouped together with agents with a certain label to play a game.

Furthermore, a constructive algorithm for establishing trust is presented. Starting with meaningless labels, agents develop preferences to interact with other agents with a certain type of label. The underlying process follows the principles of self-organization as investigated and used within Alife and Evolutionary Game Theory. So, there is no central control, agents engage only in limited local interactions, and the information exchanged among agents is partial and unreliable. Nevertheless, a stable relation of trust emerges. In addition, trust boosts the evolution of cooperation in the underlying game.

The rest of this article is structured as follows. In section two, the framework for the experiments is presented. A previously published, extended version of the Prisoner’s Dilemma allowing continuous degrees of cooperation and N-players as well as strategies for this game are shortly introduced. Section three shows how trust can be embedded into this framework. A set of labels is used to mark agents. Labels as a kind of outer appearance of agents and strategies of agents as a basis for trustworthiness are not correlated in the beginning of each experiment. An algorithm is presented, which builds trust by evolving preferences for each agent to be grouped together with agents carrying a certain label. Results based on sets of experiments are presented in section four. Experimental evidence is given that stable relations of trust can actually emerge from limited interactions among agents. In addition, it is shown that this trust boosts the evolution of cooperation, i.e., a higher cooperative level in the population is reached faster with the trust building mechanism than without it. Section five concludes the article.

## 2 The Framework for Cooperation

### 2.1 A Continuous N-player Prisoner’s Dilemma

Roberts and Sherratt published in [RS98] results on the evolution of cooperation in an extension of the standard prisoner’s dilemma (PD) to continuous degrees of cooperation. Partially inspired by this work and partially based on experiments with heterogeneous robots in an artificial ecosystem [BB98, Ste94b], following further extension to an N-player case was developed, leading to a Continuous-cooperation N-player Prisoner’s Dilemma or short CN-PD.

In the artificial ecosystem, simple mobile robots, the so-called “moles” can autonomously re-charge their batteries, thus staying operational over extended periods in time. As illustrated in figure 1, a so-called “head” can track mobile robots and it can perceive so-called pitfalls which are a kind of inverse charging stations where the batteries of the moles are partially dis-charged via a resistor. When a mobile robot approaches a pitfall, which it cannot distinguish from a charging station, the head can warn the mobile robot. The mobile robot in exchange can share the benefit of the saved energy with the head.

Let there be  $N$  moles and one head. Each mole  $m_i (1 \leq i \leq N)$  has a gain  $G_i$  based on the avoidance of pitfalls due to warnings of the head. This gain only depends on the so-called headsight  $hs \in [0, 1]$ , i.e., the percentage with which the head perceives dangerous situations. Concretely, the gain is the headsight times one energy-unit ( $EU$ ):

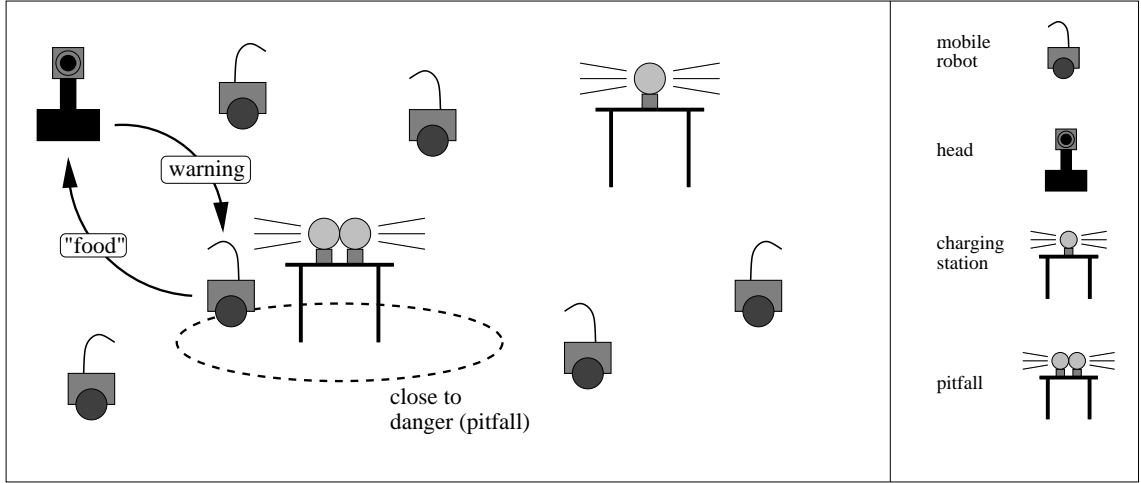


Figure 1: The extended artificial ecosystem of the VUB AI-lab, including a head and several moles. So-called pitfalls in the form of inverse charging-stations can suck energy out of a mole. Unlike moles, a head can distinguish pitfalls and charging-stations, and it can warn a mole when being close to a pitfall. The mole in return feeds a part of its benefit in form of energy to the head.

$$G_i = hs \cdot 1.0 \text{ EU}$$

Furthermore, in the beginning of each time step  $t$ , each mole  $m_i$  invests up to 0.75 energy units to feed the head. This investment  $I_i$  is proportional to the continuous cooperation level  $co_i \in [0, 1]$  of  $m_i$ :

$$I_i = co_i \cdot 0.75 \text{ EU}$$

The headsight  $hs$  depends on the amount of food the head receives from the moles, i.e., the head is completely fed when it receives the 0.75 energy units from every mole. Concretely, the headsight  $hs$  is defined as the averaged sum of cooperation levels in time step  $t$ :

$$hs = \sum_{1 \leq i \leq N} co_i / N$$

The pay-off  $po_i$  for a mole  $m_i$  is the difference between gain and investment:

$$po_i = G_i - I_i = \sum_{1 \leq j \leq N} co_j / N \cdot 1.0 \text{ EU} - co_i \cdot 0.75 \text{ EU} \quad (1)$$

So, a dilemma for the moles arises. On the one hand, it is in the interest of each mole that the head is well fed. On the other hand, there is the temptation to leave the task of actual feeding to others, as the head does not react to the behavior of a single mole, i.e., it does not punish a mole when it does not donate energy.

Note, that the pay-off for a mole depends on its own cooperation level and on the cooperation levels of all other moles. Let  $\bar{c}o$  denote the average cooperation level of the group, i.e.,

$$\bar{c}o = \sum_{1 \leq i \leq N} c o_i / N$$

The pay-off for a mole  $m_i$  can directly be computed from  $c o_i$  and  $\bar{c}o$ . Namely, the pay-off function  $f_p : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$  is

$$f_p(c o_i, \bar{c}o) = c o_i \cdot -0.75 \text{ EU} + \bar{c}o \cdot 1.0 \text{ EU} \quad (2)$$

Based on this, we can extend the terminology for pay-off values in the standard prisoner's dilemma, with pay-off types for cooperation (C), punishment (P), temptation (T), and sucking (S), as follows:

- Full cooperation as all fully invest:  $C_{all} = f_p(1.0, 1.0) = 0.25$
- All punished as nobody invests:  $P_{all} = f_p(0.0, 0.0) = 0.0$
- Maximum temptation:  $T_{max} = f_p(0.0, \frac{N-1}{N}) \geq 0.5$
- Maximum sucking:  $S_{max} = f_p(0.0, \frac{1}{N}) \leq -0.25$

For  $c o, \bar{c}o \neq 0.0, 1.0$ , we get the following additional types of pay-offs, the so-called partial temptation, the weak cooperation, the single punishment, and the partial sucking. They are not constants (for a fixed  $N$ ) like the previous ones, but actual functions in  $(c o, \bar{c}o)$ . Concretely, they are sub-functions of  $f_p(c o, \bar{c}o)$ , operating on sub-spaces defined by relations of  $c o$  in respect to  $\bar{c}o$  (table 1).

$c o < \bar{c}o$	$\bar{c}o \leq c o < 4/3 \cdot \bar{c}o$	$c o = 4/3 \cdot \bar{c}o$	$c o > 4/3 \cdot \bar{c}o$
$T_{partial}(c o, \bar{c}o)$	$C_{weak}(c o, \bar{c}o)$	$P_{single}(c o, \bar{c}o)$	$S_{partial}(c o, \bar{c}o)$
$\in ]0, 1.0[$	$\in ]0, 0.25[$	$= 0$	$\in ]-1.0, 0[$
partial temptation	weak cooperation	single punish	partial sucking

Table 1: Additional pay-off types in the CN-PD

Note that for a fixed average cooperation level  $\bar{c}o$  and two individual cooperation levels  $c o' > c o''$ , it always holds that  $f_p(c o', \bar{c}o) < f_p(c o'', \bar{c}o)$ . Therefore it holds for an individual player in a single game that:

- The partial temptation pays always better than weak cooperation.
- The partial temptation increases with decreasing individual cooperation.
- The absolute value of partial sucking increases with increasing individual cooperation.

This can also be stated as:

$$\begin{aligned} T_{max} > T_{partial}(\cdot) > C_{all} > C_{weak}(\cdot) > 0.0 \\ P_{single}(\cdot) = P_{all} &= 0.0 \\ S_{max} < S_{partial}(\cdot) < 0.0 \end{aligned} \quad (3)$$

The equation 3 illustrates the motivation for the names of the different types of pay-off. The attribute *max* for temptation  $T$  and sucking  $S$  indicates that these are the maximum absolute values. The *partial* accordingly indicates that these values are only partially reached through the related  $T$  or  $S$  functions. The attribute *weak* for the cooperation function  $C$  relates to the fact that though the player receives a positive pay-off, it is less than in the maximum cooperation case where *all* players fully cooperate. When no player invests, *all* are punished with a Zero pay-off. Whereas in the *single* case, at least the individual player we are looking at gets punished with a Zero pay-off, other players can receive all types of pay-off.

## 2.2 The Evolution of Cooperation in the Iterated CN-PD

Much like in the standard Prisoner's Dilemma, rational agents are also trapped in the global punishment as nobody will feed the head in a single game of the CN-PD. But when iterating the game over several time-steps  $t$ , strategies taking previous cooperation or non-cooperations of others into account can lead the agents into cooperation. In [Bir99], where also the Continuous N-player Prisoner's Dilemma is described in more detail, a novel strategy, the so-called Justified-Snobism (JS), is presented.

JS cooperates slightly more than the average cooperation level of the group of  $N$  players if a non-negative pay-off was achieved in the previous iteration, and it cooperates exactly at the previous average cooperation level of the group otherwise. So, JS tries to be slightly more cooperative than the average. This leads to the name for this strategy as the snobbish belief to be "better" (in terms of altruism) than the average of the group is somehow justified for players which use this strategy.

It can be shown that JS is a successful strategy for the CN-PD and especially that JS is evolutionary stable. In the experiments reported here, JS has to compete with following strategies in iterated CN-PDs:

**Follow-the-masses (FTM)** : match the average cooperation level from the previous iteration, i.e.,  $co_i[t] = \bar{co}[t - 1]$

**Hide-in-the-masses (HIM)** : subtract a small constant  $c$  from the average cooperation level, i.e.,  $co_i[t] = \bar{co}[t - 1] - c$

**Occasional-short-changed-JS (OSC-JS)** : a slight variation of JS, where occasionally the small constant  $c$  is subtracted from the JS-investment

**Occasional-cheating-JS (OC-JS)** : an other slight variation of JS, where occasionally nothing is invested

**Challenge-the-masses (CTM)** : Zero cooperation when the previous average cooperation is below one's one cooperation level, a constant cooperation level  $c'$  otherwise, i.e.,

- $co_i[t - 1] \geq \bar{co} : co_i[t] = c'$
- $co_i[t - 1] < \bar{co} : co_i[t] = 0$

**Non-altruism (NA)** : always completely defect, i.e.,  $co_i[t] = 0$

**Anything-will-do (AWD)** : always cooperate at a fixed level, i.e.,  $co_i[t] = c'$

The strategies compete in an evolutionary tournament proceeding in time-steps  $t$ . In the beginning, a population of 1000 agents is randomly created, i.e., each agent gets a randomly selected strategy following an even distribution. In the iterations  $t \rightarrow t + 1$ , the population is divided into groups with 20 agents each in a random manner. Each group plays a CN-PD. The so-called fitness  $fit(a_i)$  of agent  $a_i$  in time-step  $t$  is determined by the running average of its pay-offs, i.e.:

$$fit(a_i)[t] = \begin{array}{r} (1 - q) \cdot po_i[t] \\ + \quad q \quad \cdot \quad fit(a_i)[t - 1] \end{array}$$

with  $q \in ]0.0, 1.0[$

Reproduction of agents is proportional to their fitness as roulette-wheel selection keeps the population size fixed to 1000. In addition, new agents are randomly created in each time-step with a small likelihood  $p_{new} = 0.05$ . When running the evolutionary tournament without trust, JS multiplies and starts to take over the population (figure 5).

### 3 The Evolution of Trust

#### 3.1 The Outer Appearance of Agents

As mentioned in the introduction, trust is here seen as a kind of subjective criterion guiding the interaction with others. More concretely, a strategy is based on the objective measures on the performance of other agents, namely their cooperation-level in previous iterations. Furthermore, the given strategy  $strat_A$  of an agent  $a_A$  establishes its trustworthiness in an objective manner. If an other agent  $a_B$  would explicitly know  $strat_A$ , then  $a_B$  could rationally decide whether it is desirable to play a game with  $a_A$  or not.

Trust in contrast is based on secondary, derived measures, here the “outer appearance” of an agent in form of a marker. The main idea is as follows. Agents are randomly marked with labels from a finite set  $S_L = \{l_1, \dots, l_k\}$ . The function  $L$  maps labels to agents. Whenever a new agent  $a$  is created, a label  $l$  is randomly selected and assigned to  $a$ , i.e.,  $L(a) = l$ . Note, that this assignment is completely independent of the strategy of the agent, even during the course of the evolution. The set  $S_{color} = \{\text{red}, \text{green}, \text{blue}\}$  will serve as an extremely simple example of labels in the remainder of this article.

The co-evolution of strategies and trust proceeds in time-steps  $t$ , much like the evolution of strategies on its own. Figure 2 illustrates the overall program in pseudo-code. The crucial change to the mere evolution of strategies is that groups are not randomly formed anymore, but based on preferences on labels. The exact mechanism is described in detail in the next section.

#### 3.2 Trust as Preferences in the Group Formation

Trust as estimation of trustworthiness is established through preferences in the group-formation, i.e., agents prefer to play games with agents carrying a certain marker. Note, that this trust can not be justified by rational means alone, at least in the beginning of the iterated games, as there is no correlation between a certain label and a certain strategy.

```

1  co-evolve strategies and trust {
2  /* random initialization */
3      t = 1
4      pop[1] = ∅
5      while #pop[1] < N {
6          random create agent a
7          pop[1] = pop[1] ∪ {a}
8      }
9  /* evolutionary step (t → t + 1) */
10     while(True) {
11         ∀1 ≤ i ≤ NG : form group Gi[t]
12         ∀1 ≤ i ≤ NG : play k CN-PD on Gi
13         pop[t] = ∪1 ≤ i ≤ NG Gi[t]
14         evolve strategies on pop[t]
15         t = t + 1
16     }
17 }

1  random create agent a {
2      l = random select(SL)
3      L(a) = l
4      strat = random select(Sstrat)
5      Strat(a) = strat
6  }
7  }

```

with

- $N_G$  is the (fixed) number of groups in the population and  $N$  is the (fixed) total number of agents in the population
- `random select ( set  $S$  )` returns an element from  $S$  following an even distribution of probabilities

Figure 2: The pseudo-code of the overall co-evolution of trust and strategies.



Concretely, the so-called trust function  $trust_i : L \rightarrow [0.0, 1.0]$  of an agent  $a_i$  maps a weight  $w$  to each possible label  $l_j$ , such that  $trust_i(l_j) = w$ . The weight  $w$  represents  $a_i$ 's preference to interact with an agent with label  $l_j$ . If  $w$  is high, i.e., close to 1.0,  $a_i$  prefers to interact with agents with label  $l_j$ , or it simply trusts them. If  $w$  is low, i.e., close to 0.0,  $a_i$  prefers not to interact with agents with label  $l_j$ , or it simply does not trust them.

```

1  form group G {
2  /* initialize the group G with one agent based on fitness */
3      G = ∅
4      a = roulette-wheel selection (pop[t - 1], fit())
5      G = G ∪ {a}
6  /* add agents to G based on the trust of the agents already in G */
7      while #G < NA
8          ∀li ∈ SL : sw(li) = ∑aj ∈ G trustj(li)
9          a' = roulette-wheel selection (pop[t], sw())
10         G = G ∪ {a'}
11     }
12 }
```

with

- $N_A$  is the (fixed) number of agents per group
- `roulette-wheel selection` ( set  $S$ , function  $f : S \rightarrow \mathbb{R}$  ) returns an agent  $a$  from set  $S$  with a likelihood  $prob$  proportional to  $f(a)$ , i.e.,  $prob(a) = f(a) / \sum_{a' \in S} f(a')$

Figure 3: The pseudo-code of group-formation based on the trust-functions.

The pseudo-code program in figure 3 shows how the trust-functions are concretely used to form a new group  $G$ . The simple example of color-labels  $S_{color}$  is now used to illustrate how a new group is formed. Given the trust-functions of several agents at time-step  $t - 1$  as listed in table 2. A new group  $G$  of size  $N_A = 6$  is formed as follows. First, an agent is selected from  $pop[t - 1]$  with the roulette wheel principle based on the fitness of all agents. Let us assume, agent  $a_1$  is selected.

After the first agent  $a_1$  has been added to the group  $G$ , further agents are added in iterations of the lines 8 to 10. In the first iteration, the trust-function of  $a_1$  is used to initialize the summed weights  $sw$  for each possible label  $l_i$  (line 8). Table 3 shows the results of this first iteration and for the further iterations. In general, the function  $sw()$  is used to bias the selection of the next agent which is included in the group (line 9) for each possible label. Table 4 shows the according probabilities for this first and the further iterations with the color-set example.

agent $a_i$	$trust_i(\text{red})$	$trust_i(\text{green})$	$trust_i(\text{blue})$
$a_1$	0.241	0.987	0.328
$a_2$	0.793	0.846	0.201
$a_3$	0.086	0.392	0.003
$a_4$	0.393	0.586	0.245
$a_5$	0.230	0.567	0.045
$a_6$	0.187	0.793	0.627

Table 2: An example of trust-functions for the agents  $a_1$  to  $a_6$  in time-step  $t - 1$ .

iteration	group $G$	$sw(\text{red})$	$sw(\text{green})$	$sw(\text{blue})$
1	$G = \{a_1\}$	0.241	0.987	0.328
2	$G = \{a_1, a_2\}$	1.034	1.833	0.529
3	$G = \{a_1, \dots, a_3\}$	1.120	2.225	0.532
4	$G = \{a_1, \dots, a_4\}$	1.513	2.811	0.777
5	$G = \{a_1, \dots, a_5\}$	1.743	3.378	0.822
6	$G = \{a_1, \dots, a_6\}$	1.930	4.171	1.449

Table 3: The weights for the roulette-wheel selection of additional agents for group  $G$ .

iteration	group $G$	$prob(\text{red})$	$prob(\text{green})$	$prob(\text{blue})$
1	$G = \{a_1\}$	0.155	0.634	0.210
2	$G = \{a_1, a_2\}$	0.304	0.540	0.156
3	$G = \{a_1, \dots, a_3\}$	0.289	0.574	0.137
4	$G = \{a_1, \dots, a_4\}$	0.297	0.551	0.152
5	$G = \{a_1, \dots, a_5\}$	0.293	0.568	0.138
6	$G = \{a_1, \dots, a_6\}$	0.255	0.552	0.191

Table 4: The development of the probabilities of the color of the agent which is added to the group.

As the number of labels is usually much smaller than the size of the population, the roulette-wheel selection of an agent biased with label-preferences is done for efficiency reasons as follows. First, a label  $l$  is chosen with roulette-wheel selection using the bias  $sw()$ . Then, the population is searched in sequential order starting from a random position. The first agent with label  $l$  which is encountered is added to the group  $G$ .

Back to the color-set example, let us assume that agent  $a_2$  is selected and added to the group at the end of the first iteration. In the next, second iteration of lines 8 to 10, the summed preferences of agent  $a_1$  and  $a_2$  are used to select the next agent  $a_3$ , and so on.

### 3.3 The Update of the Trust-Functions

The trust-function of an agent  $a_i$  is updated in each time-step  $t$  based on the (very limited) experiences with other agents with a certain label. Concretely, the weight of trusting agents with label  $l_j$  is updated in each game proportionally to the pay-off and the number of agents with that label in the group. This means that when many agents with label  $l_j$  are in the group and the pay-off is high, then the agent  $a_i$  increases its trust in agents with that label  $l_j$ . A running average is used to sum the updates over consecutive time-steps:

$$trust_i(l_j)[t] = \begin{matrix} (1 - q) & \cdot & trust_i(l_j)[t - 1] \\ + & q & \cdot & po_i[t - 1] \cdot \#\{a_k \in G \text{ with } L(a_k) = l_j\} / N_A \end{matrix}$$

with  $q \in ]0.0, 1.0[$

The constant  $q$  is set to 0.1 in all experiments reported here.

agent	label	pay-off
$a_i$	$L(a_i)$	$po_i[t - 1]$
$a_1$	blue	0.257
$a_2$	green	-0.035
$a_3$	blue	0.392
$a_4$	red	0.404
$a_5$	blue	0.157
$a_6$	blue	0.289

Table 5: The (fixed) labels of the agents  $a_1$  to  $a_6$  and their pay-offs in time-step  $t$ .

Again, let us return to the example with the set  $S_{color}$  of color labels. Assume that the group  $G = \{a_1, \dots, a_6\}$  has played a CN-PD game in time-step  $t - 1$ . The pay-offs for each agent in this game and the (fixed) labels of each agent are shown in table 5. Agent  $a_4$  for example has received a rather high pay-off. As there are rather many blue agents in the group,  $a_4$  increases its trust in this color as

$$\begin{aligned}
trust_4(\text{blue})[t] &= (1 - 0.1) \cdot trust_4(\text{blue})[t - 1] \\
&+ 0.1 \cdot po_4[t - 1] \cdot \#\{a_k \in G \text{ with } L(a_k) = \text{blue}\} / N_A \\
&= 0.9 \cdot 0.245 \\
&+ 0.1 \cdot 0.404 \cdot 4/6 \\
&= 0.247
\end{aligned}$$

Note that each of the `blue` agents can have a different strategy. Especially in the beginning of the evolution, where labels and strategies are independently distributed among randomly created agents, this is very likely. Note also that the relatively high pay-off for agent  $a_4$  can be due to an exploitation of the `green` agent  $a_2$  and rather independent from the presence of the four `blue` agents.

## 4 Results

### 4.1 Trust Becomes Stable

There is neither a meaningful form of trust nor a (obvious) basis for it in the beginning of each experiment. Or more concretely, neither the labels nor the trust-functions contain any information or meaning in the beginning of each experiment:

- The labels are randomly assigned to the agents. Therefore, there is no meaningful relation between an agent’s label and its strategy.
- The trust-functions of the agents are randomly initialized. Therefore, there is no a priori, global preference of agents to be grouped together.

Nevertheless, a stable relation of trust emerges. This means, the agents evolve fixed preferences for interacting with agents with a certain label. Figure 4 shows the percentage of agents which can not “decide” which type of agent they should trust. More precisely, the graph shows the percentage of agents where the highest preference of a particular agent for a certain label in the current step is different from its highest preference in the previous step. In the beginning of the run, the percentage of “undecided” agents is very high, i.e., the agents are more or less randomly guessing in each step which type of agents they should trust. After a while, this indecision is dropping to almost Zero, i.e., the agents evolve fixed preferences for certain labels.

Note, that the basis for this evolving trust as fixed preferences is really subjective in some sense. First, it is grounded on very limited data, i.e., there are many agents with label  $l_j$  in the population, but an agent  $a_i$  builds up some belief by interacting with just a few of them. Second, within the group to which agent  $a_i$  belongs to at a time-step  $t$ , there are (most probably) many different agents in respect to labels. The update of trust does not distinguish between those labels, though different agents, and accordingly labels, do (most probably) contribute very differently to the pay-off that  $a_i$  receives.

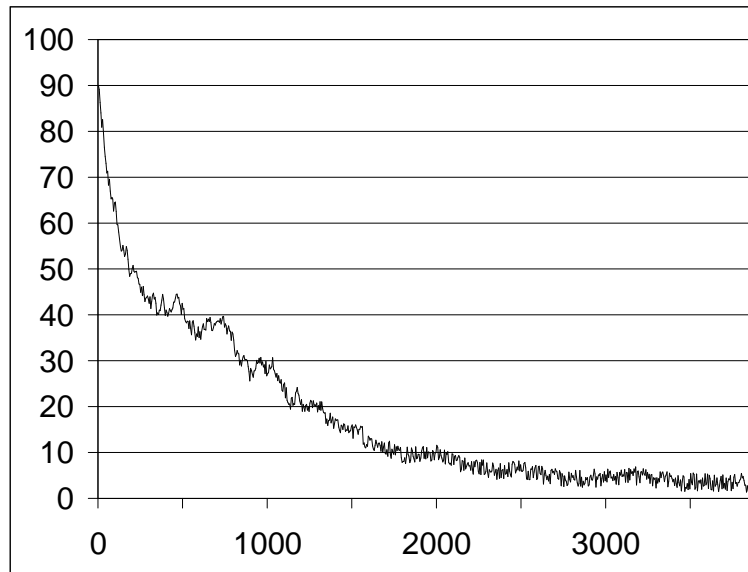


Figure 4: The percentage of agents in the population which can not “decide” which types of agents they should trust. In the beginning of the run, this percentage is high as most agents change their preference in every time step, more or less randomly guessing. After a while, fixed preferences evolve.

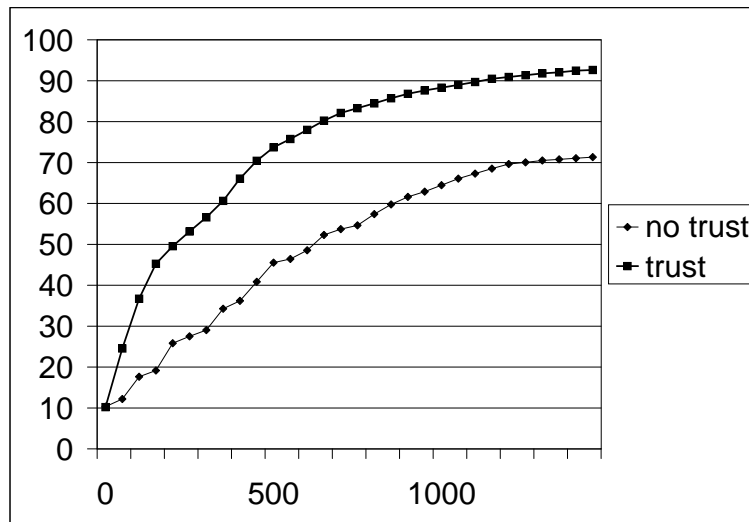


Figure 5: The general cooperation levels, averaged from respectively fifty runs with and without a co-evolution of trust. When trust is activated, a higher general cooperation level is reached much faster than without trust.

## 4.2 Evolution of Trust Boosts the Evolution of Cooperation

Despite the lack of meaning for the labels in the beginning, the evolution of trust boosts the evolution of cooperation in these experiments. Figure 5 shows the development of the general cooperation level for both cases, namely respectively fifty averaged runs with and without a co-evolution of trust. When the co-evolution of trust is activated, a higher general level of cooperation is reached much faster than without an evolution of trust.

In each of the fifty runs, the population evolved into a set of agents, which in their majority had the following properties:

- they all follow the cooperative strategy JS
- they all are marked with the same label  $l$
- they all have a high trust in the label  $l$
- they all have a low trust in other labels

This result indicates a possible explanation for the boosting of cooperation in the reported experiments. In the beginning, labels are even-distributed over agents and thus strategies. Also, preferences are even-distributed. Assume random fluctuations cause a slightly above average likelihood that trustworthy agents, which means here agents with a cooperative strategy, have a certain label. If there is in addition a slightly above average likelihood that trustworthy agents trust this label, then there is the possibility that this subtle effect reinforces itself. As a result, trustworthy agents can so-to-say recognize each other and actively group together.

## 5 Conclusion

Trust is here modeled as emergent property in a complex dynamical system. Its basis is trustworthiness, which is defined as an intrinsic property of an individual  $i_A$  in respect to an other individual  $i_B$ . It is an objective criterion in the sense that it gives  $i_B$  a measure allowing a rational choice of whether to interact with  $i_A$  or not. Unfortunately, the trustworthiness  $i_A$  is not perceivable by  $i_B$  in the general case. It is even questionable if  $i_A$  can access its very own trustworthiness for  $i_B$ , as it is an internal state which can be derived from well hidden or so-to-say unconscious processes.

The building of trust deals with the approximation of trustworthiness. When an individual meets an other one for the first time, there is no objective data from previous interactions allowing a rational choice of whether to interact or not. Subjective criteria like outer appearance must be used in those situations. Here, (in the beginning) meaningless labels are used for this purpose. Trust as approximation of trustworthiness is established through the preferences of agents to be grouped together with other agents carrying a certain marker. Groups play a game based on an extended version of the Prisoner's Dilemma. Strategies of the agents in the iterated game establish their trustworthiness.

A constructive way to update trust based on limited interactions with other agents is presented. In the experiments reported here, there is neither a correlation between labels and strategies, nor

between preferences and strategies in the beginning of each experiment. Nevertheless, stable relations of trust emerge. Furthermore, the co-evolution of trust can significantly boost the evolution of cooperation. This means that in the underlying evolutionary game, a higher cooperative level in the population is reached faster with the trust building than without it.

## Acknowledgments

Andreas Birk is a research fellow (OZM-980252) of the Flemish Institution for Applied Research (IWT).

## References

- [AH81] R. Axelrod and W. D. Hamilton. The evolution of cooperation. *Science*, 211:1390–1396, 1981.
- [Axe84] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.
- [BB98] Andreas Birk and Tony Belpaeme. A multi-agent-system based on heterogeneous robots. In Alexis Drogoul, Milind Tambe, and Toshio Fukuda, editors, *Collective Robotics, CRW'98*, LNAI 1456. Springer, 1998.
- [BG00] Micheal Bacharach and Diego Gambetta. Trust in signs. In Karen Cook, editor, *Trust and Social Structure*. Russell Sage Foundation, New York, 2000.
- [Bir99] Andreas Birk. Evolution of continuous degrees of cooperation in an n-player iterated prisoner's dilemma. Technical Report under review, Vrije Universiteit Brussel, AI-Laboratory, 1999.
- [CW94] C. Castelfranchi and E. Wemer, editors. *Artificial Social Systems – Selected Papers from the Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW-92*, number 830 in Lecture Notes in Artificial Intelligence. Springer-Verlag: Heidelberg, Germany, 1994.
- [DM91] Y. Demazeau and J.-P. Müller, editors. *Decentralized AI 2 — Proceedings of the Second European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW-90)*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1991.
- [GB99] Francisco J. Garijo and Magnus Boman, editors. *Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World : Multi-Agent System Engineering (MAAMAW-99)*, volume 1647 of *LNAI*, Berlin, June 30–July 2 1999. Springer.
- [Har98] William S. Harbison. Delegating trust. In *Security Protocols, The Sixth International Workshop IWSP'98*, LNCS. Springer, 1998.
- [JW98a] Nicolas R. Jennings and Micheal R. Wooldridge, editors. *Agent Technology; Foundations, Applications, and Markets*. Springer, 1998.

- [JW98b] Nicolas R. Jennings and Micheal R. Wooldridge. Applications of intelligent agents. In Nicolas R. Jennings and Micheal R. Wooldridge, editors, *Agent Technology; Foundations, Applications, and Markets*. Springer, 1998.
- [Lan89] Christopher C. Langton. Artificial life. In Christopher G. Langton, editor, *Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems (ALIFE '87)*, volume 6 of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 1–48, Redwood City, CA, USA, September 1989. Addison-Wesley.
- [LN98] I. Lehti and P. Nikander. Certifying trust. In *PKC: International Workshop on Practice and Theory in Public Key Cryptography*, LNCS. Springer, 1998.
- [LTFR90] Christopher G. Langton, Charles E. Taylor, J. Doyne Farmer, and Steen Rasmussen, editors. *Artificial Life II*. Addison-Wesley, Reading, MA, 1990.
- [MJW96] Jörg P. Müller, Nicolas R. Jennings, and Micheal R. Wooldridge, editors. *Intelligent Agents III; Proc. of the ECAI'96 Workshop on Agent Theories, Architectures, and Languages*. Springer, 1996.
- [MSR99] Jörg Müller, Munindar P. Singh, and Anand S. Rao, editors. *Proceedings of the 5th International Workshop on Intelligent Agents V : Agent Theories, Architectures, and Languages (ATAL-98)*, volume 1555 of *LNAI*, Berlin, July 04–07 1999. Springer.
- [Phi97] David J. Phillips. Cryptography, secrets, and the structuring of trust. In Philip E. Agre and Marc Rotenberg, editors, *Technology and Privacy: The New Landscape*. The MIT Press, 1997.
- [RS98] Gilbert Roberts and Thomas N. Sherratt. Development of cooperative relationships through increasing investment. *Nature*, 394 (July):175–179, 1998.
- [Smi84] J. Maynard Smith. The evolution of animal intelligence. In C. Hookway, editor, *Minds, Machines and Evolution*. Cambridge University Press, 1984.
- [SP73] J. Maynard Smith and G.R. Price. The logic of animal conflict. *Nature*, 246:441–443, 1973.
- [SRW98] Munindar P. Singh, Anand Rao, and Michael J. Wooldridge, editors. *Proceedings of the 4th International Workshop on Agent Theories, Architectures, and Languages (ATAL-97)*, volume 1365 of *LNAI*, Berlin, July 24–26 1998. Springer.
- [Ste94a] Luc Steels. The artificial life roots of artificial intelligence. *Artificial Life Journal*, 1(1), 1994.
- [Ste94b] Luc Steels. A case study in the behavior-oriented design of autonomous agents. In Dave Cliff, Philip Husbands, Jean-Arcady Meyer, and Stewart W. Wilson, editors, *From Animals to Animats 3. Proc. of the Third International Conference on Simulation of Adaptive Behavior*. The MIT Press/Bradford Books, Cambridge, 1994.
- [WMT96] M. Wooldridge, J. P. Muller, and M. Tambe, editors. *Intelligent Agents Volume II—Proceedings of the 1995 Workshop on Agent Theories, Architectures, and Languages (ATAL-95)*. Springer-Verlag, Berlin, 1996.