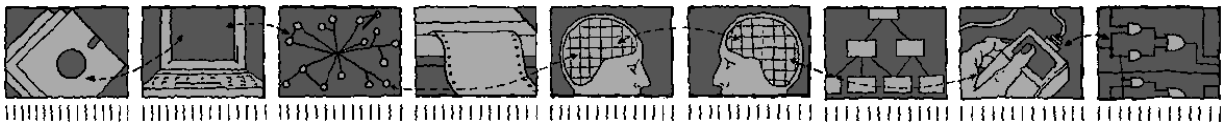
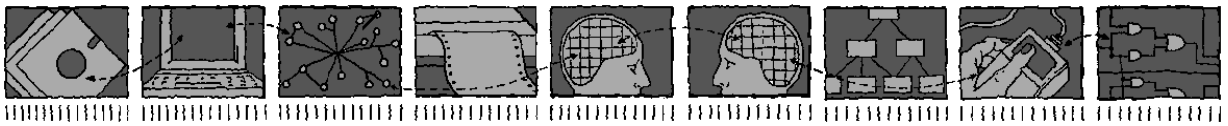


Department of Computing Science and Mathematics
University of Stirling



Optimism and Pessimism in Trust

Stephen P. Marsh

Technical Report CSM-117

August 1994

Department of Computing Science and Mathematics
University of Stirling

Optimism and Pessimism in Trust

Stephen P. Marsh

Department of Computing Science and Mathematics, University of Stirling
Stirling FK9 4LA, Scotland

Telephone +44-786-467444, Facsimile +44-786-464551
Email spm@cs.stir.ac.uk

Technical Report CSM-117

August 1994

Abstract

Artificial agents do not exist in the world in solitude. They are required to interact with others, and thus must reason in some fashion about those they interact with. This paper presents a view of trust which artificial agents can use to help in such reasoning processes, providing them with a means to judge possible future behaviour on the basis of past experience. The paper discusses the notion of 'dispositions' of trusting behaviour, which we call Optimism, Pessimism and Realism. Each different disposition results in different trust estimations from an agent. We discuss the possible effects of these differences. Finally, we present the concept of memory in trusting agents, and briefly suggest some ways in which memory spans can affect the trusting decisions of such agents, with different dispositions.

1 Introduction

In the course of the last three years, we have been developing a formalism for the social phenomenon of trust. We have argued previously that the inclusion of at least an understanding of trust in artificial agents will provide robustness under uncertainty, and an addition to the decision-making repertoires of that agent [9, 10]. In addition, we propose that the suggested development of such a formalism is of use in itself since it can help to provide a deeper understanding of the workings of the phenomenon, which is both vaguely defined and badly understood at present. Indeed, much of the relevant literature is either vague or not in the mainstream of its field [8].

This report addresses one of the aspects of trust that has come to light in the research that has been carried out, namely the concept of dispositions, and how they can affect the way an artificial agent makes trusting decisions, ultimately affecting the agent's final decision. We propose that such insights are also partially applicable to the human sphere. Some of the questions that can be addressed using the formalism are as follows:

1. Is it 'good' to be an optimist? [1]
2. Is it 'good' to be working with an optimist?
3. How long does it take before a pessimist or an optimist forgets? What difference does this make to their trust-based decisions?
4. What difference can such dispositions make in cooperative situations? Is it the case, for example, that optimists are 'better' people to work with than pessimists?
5. Should we be nicer to optimists than pessimists, and can either be easily exploited? This question is perhaps a little odd — clearly we do not wish to exploit agents, but an understanding of *how* they may be exploited is important in order to prevent such exploitation.

Point 3 here brings to light another of the discussion points that we present here, that of the memory span of an artificial agent. Clearly, we cannot expect agents to remember for ever — apart from being unrealistic, it is also physically impossible. They can, however, be set up to remember far more than we as humans could ever hope to, and for longer, with greater accuracy. This being the case, the memory span of an agent becomes important when discussing the concept of trust, which is inherently experiential. In other words, the trust placed in other agents is in part a function of the experiences that the trustee has had with the other agents, in similar and diverse situations [10]. If we restrict the agents' memory span, we restrict the amount of information it knows with regard to such experiences, and thus we can affect the trusting decisions the agent makes.

The development of a formalism for trust is of great utility to these discussions because:

- It brings to the fore the concept, and provides its own means for the discussion and clarification thereof.
- It raises important questions, such as those above, and goes some way toward providing the means to answering them.

1.1 Overview

In the remainder of this report, we address and seek to answer some of the questions raised above. The next section discusses in more details the concepts of dispositions, presenting an idealised view of a spectrum of dispositions, along which all agents lie. Following this, we present a discussion of the phenomenon of trust, then a summary of the formalism as it presently stands. It is worth noting here that the formalism is an article which is continually in flux. By its nature, it can

[1] The term 'good', despite its subjectivity, is a fair way of saying that the positive utility gained from a situation is greater than the negative utility associated with that situation — something is 'good' when we get something worthwhile out of it.

not be fixed, since it provides the mans for its own discussion and refinement, and at the present time there is a sparse knowledge of the actual workings of trust (see the author's thesis for further discussion of this [11]). The formalism is presented here for two reasons, then:

- It stands on its own to show that a simple formalism for the concept can be developed (and implemented — see [11]).
- It provides a means for the discussions of dispositions and memory span which follow.

The following section discusses the idea that trusting dispositions, in particular optimism, pessimism and realism, as we call them, can make a difference to both the workings and the final decisions involved in trust. It also presents a brief overview of a testbed which is being developed to test such theories.

We then present a discussion of the concept of memory in trusting agents, and how the memory span of an agent, allied to particular dispositions, is of crucial importance to the agent's final trusting decisions.

Finally, we present a brief list of conclusions and ideas for further work in the area, in particular as regards allying the concept of artificial trust with both other decision-making strategies and testing the phenomenon in less constrained environments.

2 Optimism and Pessimism

The following presents an idealised notion of the optimistic and pessimistic dispositions. It is a simplified, non-trivial, account of these dispositions, and much more detailed analysis would suggest that, for example, some optimists do trust others very little. There are two points to make here. One is that, we provide a generalisation here, such that most optimists do trust others relatively highly, and pessimists trust relatively little. Secondly, the presentation of such a simplified account is useful in itself — simplification is the key to understanding [15], since on it, we can build greater complexity, gradually approaching an identity with the phenomenon we discuss [1].

Optimism is defined as "1. the tendency to expect the best in all things. 2. hopefulness; confidence..." (Collins Dictionary, 1991). In terms of DAI, or interactions with others, this means that an optimist is one who will look for the best in those with whom they interact. We can also look for an optimist to be forgiving. In other words, speaking with relation to trust, the optimist is likely to be one whose trust in others is high, and inflexible in a downward direction. Thus, following his being exploited by another, his trust in that other will not decrease by too much.

The pessimist, however, will look upon the status of others as something to be proved. The amount of trust he has in others will be relatively inflexible in an upwards direction, and a small exploitation by another will result in drastic loss of trust, whilst continued cooperative behaviour by the other will result in only small gradual increases in trust.

In figure 1, we present a view of the spectrum of trusters between the extremes of optimism and pessimism. All agents have varying degrees of optimistic or pessimistic attitudes. Indeed, these attitudes may vary from situation to situation, or agent to agent, depending on past experience, hearsay, or the characteristics of particular situations. Particular agents may, for example, start out with an optimistic frame of mind relative to a particular other agent, only to find, after continued disappointment, that their disposition verges on the pessimistic.

Figure 1 suggests a continuum. Along this continuum we may all exist. Certainly, in the artificial world of DAI, we can categorize our agents along this line easily. That done, we can allow our reasoning agents to reason about each other along the lines discussed above. The following section presents some axioms based on the discussion above, and using the formalism given in [9, 16].

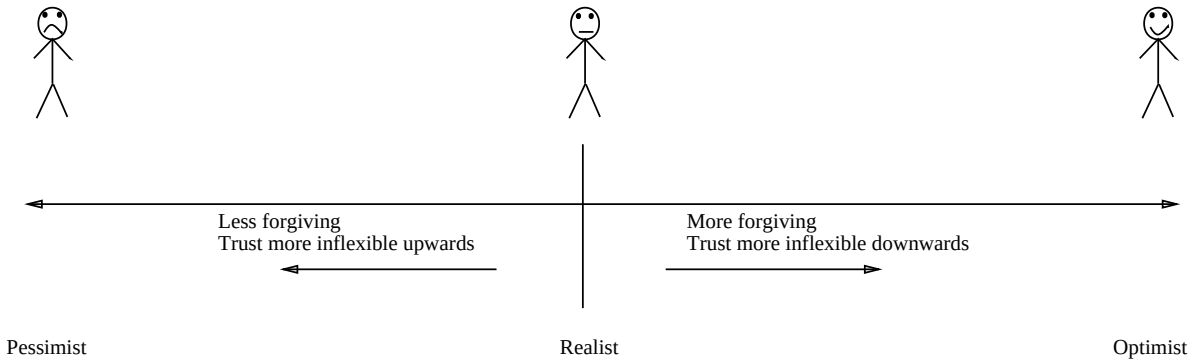


Figure 1: Spectrum of Realism

3 Why Trust

In cooperative situations, indeed in life in general, trust is a salient factor in many of our decisions [2, 7, 17]. It is surprising, then, that the phenomenon is so little understood or investigated [8]. The formalism presented in Marsh (1992) and revised in Thimbleby *et al.* (1994) goes some way towards correcting that omission. The formalism itself makes certain assumptions about how trust behaves, and ultimately presents one method for the clarification of the phenomenon of trust, albeit a relatively simple one which is easy to understand. One of the formalism's strengths is its inherent extensibility; assumptions made at the outset can be readily built into the formalism as hard rules, likewise exceptions can be easily spotted and corrected, via extensions or changes to the original formalisation.

The initial goal of the formalism was to provide artificial agents in DAI with the wherewithal to reason with and about trust. Thus it was made simple, small, and straightforward. In DAI, agents are faced with the task of working with and around others, each of which is independent, with the aim of getting a specific job done [3]. Many of these jobs require more than simple coordination; rather they require that the agents cooperate with each other. The gap between coordination and cooperation is long, and new methods of reasoning are necessary to help our agents cross it. There are several answers to this problem, one of which is to have a controlling agent which coordinates all others, thus ensuring cooperation, another is to assume that all agents are implicitly trustworthy and cooperative, and so we need not consider what happens with untrustworthy, completely uncooperative, self-interested agents [14]. The first, although attractive, loses one of the key strengths of DAI, that of graceful degradation — should the master agent fail, all the slave agents are left leaderless and 'clueless'. In the second, the dangers are obvious — the real world within which we wish our agents to operate does not provide for totally trustworthy agents, rather it provides agents of all dispositions and creeds, a world in which agents which trust blindly will quickly suffer.

Trust, then, presents itself as a way out of this predicament. In human life, indeed, it has already been accepted as a way of coping with the freedom of others to do as they wish [8, 6]. In addition, it allows us to cope with the massive complexity of everyday life [7] by trusting various things to happen or not to happen. As such, it is the ideal tool for allowing artificial agents to both reason about entering into cooperative situations with other agents, and to exist on a moment to moment basis in the complex real world. Formalising the concept allows us to incorporate it into our agents with the minimum of effort.

The formalism, as presented in Marsh (1992) [9] and Thimbleby *et al.* (1994) [16] has its own problems. Not least is the number of assumptions that have been made on the way to the final formalisation. One such concerns how the disposition of an agent can affect the final analysis of trust, in terms of whether the agent is optimistic or pessimistic, or somewhere in between. This paper clarifies this omission.

3.1 The Formalism

The formalism presented here is concerned with cooperation between two (or more) agents, from the point of view of one of those agents, x, with reference to the other, y.

We notate "the amount x trusts y" by $T_x(y)$.

$T_x(y)$ has a value in the interval $[-1, 1)$ (i.e., $-1 \leq T_x(y) < +1$); 0 means no trust; -1 represents total distrust. The two are not the same since, the situation of 'no trust' represents when the truster has little or no information about the trustee, or is indifferent. The 'total distrust' situation is a proactive measure, requiring that the truster think about what she is doing. A value of $+1$ is not allowed for the reason that 'blind trust' implies that the truster does *not* think about the situation, since, by definition, they trust blindly, giving trust without hesitation. This does not fit with our definitions of trusting behaviour. [2]

An agent is, at any time, in a *situation*. A situation can be defined, then, as a point (or several points) in time relative to a specific agent. [3] Particular situations have particular levels of importance to agents, dependent on various circumstances, all, or mostly, based in the agent. We represent x's view of the importance of a situation α by $I_x(\alpha)$. I is taken to be a scalar here, since the inherent uncertainty of a situation's outcomes prohibits the use of vectors describing every possible outcome [16].

$I_x(\alpha)$ has a value in $(-1, 1)$. Whilst this is an agent-based measure, in a human system an estimate of the importance of a particular situation may be relatively easy to ascertain. For a computer-based agent, there are many different ways of determining the importance of a situation, such as payoff functions (cf Rosenschein (1985)).

We represent the utility (cost/benefit) of situation α for x with $U_x(\alpha)$, with value in $[-1, 1]$ — we normalise utility to be in this range. It might be conventional to use 'cost' as the weight, but it is more convenient to take a value (utility) that correlates with trust. Utilities can be negative, since it is often the case (as with money), that one agent gains what another loses.

We informally define trust of an agent x in y (in some given meeting) as the probability weighted by UI that x acts to achieve any outcome *as if* it trusts y. In other words, trust is the degree of certainty that people act to increase one's utility. "I don't know what y will do, but I trust him just so-much to have my best interests at heart in his actions," is the notion captured by the more formal expression. This goes along with many presented definitions of trust — see in particular Gambetta (1990)[6].

Defining trust

For two different [4] agents, x and y, in a situation α, from x's point of view, we represent the amount of trust x has in y with the notation $T_x(y, \alpha)$. This is to be read as, "the trust of x in y in the situation α." This takes into account the fact that different situations require different levels of trust, even in the same person [9].

We now come to formulae for determining trust in a cooperative situation. The values expressed within the formulae below would ordinarily be estimated by the agent concerned. In the case of humans, values for initial trust, importance of situation, and so forth would be estimated by the human, and would most likely be different in each case, for each team member. Alternatively, we (or the computer) might estimate them by parameter fitting.

To determine situational trust, notated $T_x(y, \alpha)$:

$$T_x(y, \alpha) = \widehat{T_x(y)} U_x(\alpha) I_x(\alpha)$$

where $\widehat{T_x(y)}$ is an estimate x has of how much he can trust y. It is the calculation of this estimate which is altered depending on the disposition of the agent making the decision. The next

[2] For more on this, see the author's thesis. The distinction is in fact part of the philosophy of certain Scottish philosophers such as Adam Smith and Dugald Stewart [4].

[3] The granularity of time here is glossed over, since some situations may be of the order of seconds (e.g. my turning the radio on), and some may be far longer (my working towards a PhD, for example).

[4] Ordinarily, the two agents are not the same, and $x \neq y$. There are situations involving *one* agent trusting itself, and this aspect is discussed in [11]

section discusses a realist approach to the estimate, and is followed by a discussion of the different approaches taken by optimists and pessimists.

4 Realism

The realist approach from many angles appears to be a sensible method of attaining the estimate of trust We have in fact suggested two methods of obtaining a realist estimate: the mean and the modem (see Thimbleby *et al.*. 1994).

To find a mean value for use as the estimate, other factors have to be borne in mind. The equation suggested in Thimbleby *et al.* (1994), implied that $\widehat{T_x(y)}$ may be an average over a sample of tasks: [5]

$$\widehat{T_x(y)} = \frac{1}{|A|} \sum_{\alpha \in A} T_x(y, \alpha)$$

Here, A is a set of situations. Deciding which situations the agent can remember (see below) to include in this set is non-trivial: inclusion of some situations may result in very different decisions from the agent. There are several options:

1. We can include all situations the agent remembers.
2. We include only those situations the agent can remember where y played a part (this is in fact what the equation above suggests).
3. We can include only those situations the agent can remember where y played a part *and* which were identical (or very similar) to the present situation (here, α).

Clearly, different choices from the above may result in different estimates, and thus different final trust values and decisions. The sensible compromise is to take the second option, although the third gives a more accurate representation. There may be no choice: the agent may not actually know the other (and so must realistically choose the first option, or a permutation of it) or may know the other, but not in similar situations (and so, the second option can be chosen).

The mode is less intuitively obvious as a measure for realists in estimating trust. The trust value is in fact continuous, and it is likely that there will be no repeated exact values, simply because of the many possibilities of different situations, agents, and so forth: estimates may even differ from day to day depending on how the truster has been treated that morning, for example. We do not discuss it further here, but mention it only as a possible alternative to the use of the mean for realist estimates in trust.

4.1 Optimism and Pessimism

In Thimbleby *et al.* (1994) we suggested other measures for the value of $\widehat{T_x(y)}$, two of which are of interest here. Firstly, we suggested that, were x an optimist, she would more likely choose the maximum of trust values in A than the average. Pessimism is likewise presented. Indeed, there are two major differences between optimism and pessimism here, The first is the manner in which the final trusting value is chosen, the second concerns the amount that the value of trust is altered by in the light of experience — how much, for example, the optimist increases trust following cooperative behaviour by another, and how much the pessimist would do so. We address these questions here. In order to discuss the notions, we need to introduce temporal considerations to the formalism presented above. We us the superscript t to notate a specific instant in time, thus, for example: $T_x(y)^t$ represents the amount that x trusts y at time t.

[5] This is only over a sample in A since agents are assumed not to have unbounded memories.

4.1.1 Optimists

To summarize the above in notation. If x is an optimist, and y's disposition does not matter: [6]

$$T_x(y,\alpha) = \widehat{T_x(y)} U_x(\alpha) I_x(\alpha)$$

Where:

$$\widehat{T_x(y)} = \max_{\alpha \in A}(T_x(y,\alpha))$$

When we suggest that A is a collection of situations, we can have two interpretations. Firstly, A is all of the situations which x has been in (or can remember — see below), without exception. This leads to various problems, since, as is mentioned above and in Marsh (1992) [9], different kinds of situations require different levels of trust. Thus, taking a maximum (in this case) trust for all situations is unrealistic — it may be the case, for example, that x was in a situation some time ago which resulted in a very high payoff, thus trust there was unrealistically high. In the present situation (α), things could be very different, and similar situations in the past have convinced x of the need for a fairly low level of trust. Thus, the second option is more sensible, to take A as a sample of all the situations x has experiences, with its members only those situations which are, as far as x is concerned, significantly similar to α. Once again, this decision is x's, and is, therefore, subjective. Mistakes, then, can happen, but the resultant value of trust in that situation is likely to be closer to what it should more sensibly be. In the example used above, it would be considerably lower than the unrealistically high outcome with the large payoff.

If we represent the total life history for an agent with A_T, as a set with n members (see section 5), for example $\{0.54_a, 0.21_b, 0.25_c, 0.34_d, 0.98_e\}$, where the subscripts are simply identifiers for situations, then we can say that, for example, situations b and c were similar to situation α, that x is presently in. The resultant $\widehat{T}_x(y)$ is thus 0.25 (the value for c) for this situation.

Following a cooperative decision in a situation from x at time t, and a defection by y, there are two options for the optimist:

1. The trust x has in y does not change. Naturally, this option is more likely should the costs to x be low when y defects. The higher the costs, the more likely x will choose option 2. For this option, however:
$$T(x,y,\alpha)^{t+1} = T(x,y,\alpha)^t$$

2. The trust x has in y will decrease by some amount. The actual amount, δ_x is dependent on the cost to x of y's defection, the importance of α to x ($I_x(\alpha)$), and possibly the mount of trust there was to start with ($T_x(y,\alpha)$). It is also likely to depend on x, notably how much of an optimist he is (the more of an optimist, the less δ_x may be). Thus we append x to δ. Here, then:
$$T(x,y,\alpha)^{t+1} = T(x,y,\alpha)^t - \delta_x$$

Following a cooperative decision by x in y, and cooperative moves from y, trust may increase, or it may stay as it was:

1. In extreme cases, the amount of trust x has in y will remain static. These cases are extreme because, as a generality, trust would increase. Examples of situations where it may not are, when trust is already extremely high, or when the benefit to x is very low following y's collaboration. In these cases, x may decide not to increase his trust in y, [7] thus:
$$T(x,y,\alpha)^{t+1} = T(x,y,\alpha)^t$$

[6] This sounds strange: in fact, we are discounting any extraneous variables, such as y's possible disposition or trusting behaviour, or even whether y is trusting, along with many other environmental variables, in order to illustrate the workings of the formalism clearly.

[7] Again, this is largely a subjective matter, and may depend on situation specifics, which cannot be predicted before they occur.

2. In more normal situations, the amount of trust will increase by $\psi_{\ x}$. Again, this amount is subjective, and decided on a situation by situation basis by the agent concerned. Here, then:

$$T(x,y,\alpha)^{\,t+1} = T(x,y,\alpha)^{\,t} + \psi_x$$

For an optimist, the following condition holds:

$$\psi_x \geq \delta_x$$

Thus, the amount of increase following cooperation is generally greater than (and sometimes equal to) the amount of decrease following defection. This will become more clear following our discussion of pessimists.

4.2 Pessimists

The definitions for pessimists are similar to those for optimists. Firstly, if x is a pessimist, and y's disposition does not matter:

$$T_x(y,\alpha) = \widehat{T_x(y)} U_x(\alpha) I_x(\alpha)$$

Where:

$$\widehat{T_x(y)} = \min_{\alpha \in A}(T_x(y,\alpha))$$

Again, we take A to be the set of situations that x remembers that x considers to be similar to α. Thus, for $A_{\,T}$ as $\{0.54_{\,a}, 0.21_{\,b}, 0.25_{\,c}, 0.34_{\,d}, 0.98_{\,e}\}$ and with situations b and c similar to the current situation (α), the pessimist would select $\widehat{T}_x(y)$ to be 0.21, the value for b, as opposed to 0.25 for the optimist, above.

Following a cooperative decision in situation α for x at time t, and a subsequent defection by y, the pessimist, as the optimist, has two choices. The difference between pessimists and optimists lies in the way these decisions are made:

1. In extreme cases (for optimists, this is the normal case), the amount of trust x has in y will not change. These cases are subjective, in that x makes the decision based on situation specifics, such as how much the situation cost him, or other factors, such as how low his trust in y already is. In this case, then:

$$T(x,y,\alpha)^{\,t+1} = T(x,y,\alpha)^{\,t}$$

2. More normally, the trust x has in y will decrease, possibly by quite a large amount. We notate this $\epsilon_{\,x}$. Thus:

$$T(x,y,\alpha)^{\,t+1} = T(x,y,\alpha)^{\,t} - \epsilon_x$$

In the opposite case, where y reciprocates cooperation, the following options are available to the pessimistic x:

1. The trust x has in y may remain static. Once again, this depends on many subjective and objective decisions made by x:

$$T(x,y,\alpha)^{\,t+1} = T(x,y,\alpha)^{\,t}$$

2. There is a possibility that x will increase the amount of trust he has in y. For a pessimist, we could argue that this was unlikely. Rather than doing this, and ending up with a completely static trust (as in option 1), we suggest that the trust will increase, possibly by quite a small amount (the magnitude is, again, dependent on various unpredictables), which we notate μ_x. Thus:

$$T(x,y,\alpha)^{\,t+1} = T(x,y,\alpha)^{\,t} + \mu_x$$

For the pessimist, the following rule holds:

$$\epsilon_x \geq \mu_x$$

4.3 Discussion

The rankings for the amount of adjustment made to trust following observed behaviour of another are given for optimists and pessimists. Bringing them together, we suggest that:

$$\epsilon_x \geq \psi_x \geq \delta_x \geq \mu_x$$

Here, the ϵ_x and ψ_x could feasibly be exchanged, as could the δ_x and the μ_x, to represent a 'better' world.

For an agent at the centre of the spectrum given in figure 1, the following is true:

$$\epsilon_x = \psi_x = \delta_x = \mu_x$$

Thus, although the magnitudes of alteration are decided at the time of the situation, and dependent on several unknown variables, such as cost or benefit of situation, and so forth, for any particular situation, they are identical.

In fact, this stipulation is somewhat limiting. What is a better representation is:

$$\epsilon_x = \psi_x$$

$$\delta_x = \mu_x$$

And:

$$\epsilon_x \geq \delta_x$$

5 Memory

We turn lastly to the concept of memory. The definitions of trust given above rely on the agents' memory of situations which have come before. Optimists take the maximum value of trust in these preceding situations to calculate the value of trust for the present situation. The question arises, then, of how many situations — how far back — the agent can remember. In other words, what is the size of the set A_T.

Consider an agent with a memory of 1; that is, he remembers only the result (and the estimated trust value) of the previous interaction with a particular agent. It would not matter, then, whether he were an optimist or not — the result is identical in all cases. The longer the memory of the agent, the more such a disposition matters.

The concept of memory is more fully discussed in the author's thesis. Briefly though, some problems remain — in optimism and pessimism, a memory size of which is anything short of unbounded may at some time result in erratic behaviour due to peaks or troughs in the values of $T_x(y)$ at some time in the past. Consider a peak value which occurred several time periods ago. If that peak is in the history of the optimist, it will always be chosen to substitute for $\widehat{T_x(y)}$. Once it passed out of the agent's memory range, it would not be taken into account, and another peak will be chosen. This may be considerably less than the original, resulting in a much changed value for situational trust from one situation to the next, even for 'identical' situations. That said, such a situation should not come about since we expect trust to be a gradually changing value over time, with only slight changes from one situation to the next. Using a simple average, however, would rid us of such considerations.

6 The Testbed

During the course of this work, the need for an implementation of trust became clear. It is of use for two major reasons:

- It provides a clear view of the state of the formalism — both in terms of the behaviour and decisions of trusting agents and in terms of the applicability of the formalism to implementation.

- It provides an additional justification in that the formalism can be seen to be working in artificial agents.

A preliminary testbed has been designed and implemented in HyperCard [8] on a Macintosh. It consists of a 'PlayGround' which is a grid populated by several independent agents. Each agent may be a trusting entity, or may be a random cooperator/defector. Agents are free to wander around the PlayGround until they meet another, and are then put into a forced Prisoners' Dilemma. All agents are referred to by name (which is a letter here) and all payoffs, costs and benefits are known. In addition, agents can have variable memory lengths, which can be set by the user. Finally, each trusting agent can have one of the three dispositions, optimist, pessimist, realist, and each follows the general rules for that disposition given above. A diagram of one state in the testbed is given in figure 2.

Figure 2: A simple testbed for determining trusting agents' behaviour.

The testbed fulfills the following criteria, amongst others:

1. It provides the user with control over the positions of particular agents.
2. It provides a limited concept of society, with several agents present at any one time.
3. As in 'real' societies, each agent has a fair chance of encountering many others through a certain time period. Thus one of the problems of the Iterated Prisoners' Dilemma is removed — that of the falseness of being forced to interact with one agent over and over.
4. Agents are able to influence their movement in favour of a particular direction (in the simple case, they can move towards those they trust and away from those they do not). This again provides freedom of choice, but constrains it since they can only influence their direction, never choose it exactly — this strategy was chosen to reflect the idea that in the future we

[8] HyperCard is a trademark of Claris Corporation.

wish agents to work for *us*, not themselves. This being the case, they may be forced to go where they do not with to go in order to get a job done. This is, in fact, not too far removed from human activity — some go to work because they *have* to, rather than because they *want* to.

5. The user has access to agent specific details, such as basic and general trust, costs and benefits of situations, and so forth, and can change these at will. It is also possible to access directly the 'memory' of an agent and change aspects of this also.

6. Results can be exported and analysed in detail.

Further details of the implementation can be found in the author's thesis [11]. A more detailed, collaborative implementation of trusting agents, based on a game of negotiation and strategy, and including other decision making and belief strategies, is at its design stage.

Several experiments have been performed using the testbed (see [11]), and the results have been promising. With particular relevance to this report are the following findings:

1. An optimist with a high [9] trust can 'educate' cooperation from a trusting agent of any disposition, providing that:

   (a) The trust of the other increases following cooperation by the first.

   (b) The trust of the other increases to above the cooperation threshold for that agent *before* the optimist's trust decreases below it's cooperation threshold.

2. This finding can be duplicated with pessimists educating cooperation, but they are more likely to decrease their trust to below their cooperation threshold first.

3. The shorter the memory span, the less disposition matters — agents with a memory of 0 are in effect random cooperators/defectors. In other words, memory *does* matter.

4. Following from this, the longer the memory span, the more disposition matters, especially from the point of view of extreme dispositions (optimism and pessimism) and extreme values. For example, an optimist with an extremely high initial trust will continue trusting and cooperating long after a reasonable person would have given up, consequently losing a great deal. This is an example of pathological trusting, and can exist in humans [5].

5. We found no real advantage to being optimist or pessimist — indeed, they showed extremes of trusting or non-trusting behaviour which were at times disadvantageous.

6. A memory span of about 10 iterations with any one agent appeared to be satisfactory — this is a useful finding since it shows that experiential trust can be incorporated into agents where space is a problem.

For more details, see the author's thesis.

7 Conclusions and Ideas for Further Work

We have briefly presented a formalisation of the dispositions of optimism and pessimism in trust. Formal descriptions of optimism were presented, from which descriptions of pessimism can easily be deduced. We touched on the problems that a limited memory span in agents might cause, particularly if the agent concerned uses extremes to determine situational trust. The formalism presented in Marsh (1994) and summarised here has been used throughout in order to provide a clear and concise discussion of the various concepts involved, and has proved useful in generating insights into the behaviour of trusting agents. To answer one of the questions presented in the

[9] The question of values is problematic. However, 'high' in this sense is fairly self-explanatory, but can be taken as, for example, above 0.8 for an agent.

introduction: an optimist with a very long memory is better to work with than a pessimist, but the shorter the pessimist's memory span, the less it matters.

Development of the formalisation for trust is not yet complete. Indeed, one of its strengths is that it will remain 'unfinished' [12, 13]. In addition, it provides the mans for its own discussion and refinement. It is therefore a practical tool for the social sciences. The brief presentation of optimism and pessimism in this paper is an indication of the capabilities of the formalism with regard to the precise discussion of trust and associated concepts.

References

[1] Birkhoff, George David. 1956. A Mathematical Approach to Ethics. *Pages 2198 – 2208 of:* Newman, James R. (ed), *The World of Mathematics, Volume 4.* New York: Simon and Schuster.

[2] Bok, Sissela. 1978. *Lying: Moral Choice in Public and Private Life.* New York: Pantheon Books.

[3] Bond, Alan H., & Gasser, Les. 1988. An Analysis of Problems and Research in DAI. *Pages 3–35 of:* Bond, Alan H., & Gasser, Les (eds), *Readings in DAI.* California: Morgan Kaufmann.

[4] Broadie, Alexander. 1991. Trust. Presentation given for the Henry Duncan prize, the Royal Society of Edinburgh, 2nd December.

[5] Deutsch, Morton. 1973. *The Resolution of Conflict.* New Haven and London: Yale University Press.

[6] Gambetta, Diego (ed). 1990. *Trust.* Oxford: Basil Blackwell.

[7] Luhmann, Niklas. 1979. *Trust and Power.* Chichester: Wiley.

[8] Luhmann, Niklas. 1990. Familiarity, Confidence, Trust: Problems and Alternatives. *Chap. 6, pages 94–107 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

[9] Marsh, Stephen. 1992. Trust and Reliance in Multi-Agent Systems: A Preliminary Report. *In: MAAMAW'92, 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Rome.*

[10] Marsh, Stephen. 1994. *Trust in DAI.* To appear, Springer LNAI, June 1994.

[11] Marsh, Stephen. In preparation, 1994. *Formalising Trust as a Computational Concept.* Ph.D. thesis, Department of Computing Science, University of Stirling.

[12] Popper, Karl R. 1967. *The Logic of Scientific Discovery.* Hutchinson, London.

[13] Popper, Karl R. 1969. *Conjectures and Refutations.* Routledge and Kegan Paul, London.

[14] Rosenschein, Jeffrey S. 1985. *Rational Interaction: Cooperation among Intelligent Agents.* Ph.D. thesis, Stanford University.

[15] Simon, Herbert A. 1981. *The Sciences of the Artificial (Second Edition).* MIT Press.

[16] Thimbleby, Harold, Marsh, Steve, Jones, Steve, & Cockburn, Andy. 1994. Trust in CSCW. *In:* Scrivener, Steve (ed), *Computer Supported Cooperative Work.* Ashgate Publishing.

[17] Yamamoto, Yutaka. 1990. A Morality Based on Trust: Some Reflections on Japanese Morality. *Philosophy East and West*, **XL**(4), 451–469.