Formalising Trust as a Computational Concept

Stephen Paul Marsh

Department of Computing Science and Mathematics

University of Stirling

Submitted in partial fulfilment of
the degree of Doctor of Philosophy

April 1994 1/95

Abstract

Trust is a judgement of unquestionable utility — as humans we use it every day of our lives. However, trust has suffered from an imperfect understanding, a plethora of definitions, and informal use in the literature and in everyday life. It is common to say "I trust you," but what does that mean?

This thesis provides a clarification of trust. We present a formalism for trust which provides us with a tool for precise discussion. The formalism is implementable: it can be embedded in an artificial agent, enabling the agent to make trust-based decisions. Its applicability in the domain of Distributed Artificial Intelligence (DAI) is raised. The thesis presents a testbed populated by simple trusting agents which substantiates the utility of the formalism.

The formalism provides a step in the direction of a proper understanding and definition of human trust. A contribution of the thesis is its detailed exploration of the possibilities of future work in the area.

Summary

1. Overview

This thesis presents an overview of trust as a social phenomenon and discusses it formally. It argues that trust is:

- A means for understanding and adapting to the complexity of the environment.
- A means of providing added robustness to independent agents.
- A useful judgement in the light of experience of the behaviour of others.
- Applicable to inanimate others.

The thesis argues these points from the point of view of artificial agents. Trust in an artificial agent is a means of providing an additional tool for the consideration of other agents and the environment in which it exists. Moreover, a formalisation of trust enables the embedding of the concept into an artificial agent. This has been done, and is documented in the thesis.

2. Exposition

There are places in the thesis where it is necessary to give a broad outline before going deeper. In consequence it may seem that the subject is not receiving a thorough treatment, or that too much is being discussed at one time! (This is particularly apparent in the first and second chapters.) To present a thorough understanding of trust, we have proceeded breadth first in the introductory chapters. Chapter 3 expands, depth first, presenting critical views of established researchers.

3. Formalism

Formalisation of trust shows that it can be grasped in a simple fashion using simple mathematics. It is revisable. It provides the tools necessary for its own revision and discussion. It provides the basis for trusting artificial agents, which could form stable coalitions, take 'knowledgeable' risks, and make robust decisions in complex environments. The thesis presents experimental results from simple implementations of artificial trusting agents, and proposes several avenues for further research.

4. Contribution to the Area

Although the subject of trust has been widely discussed, particularly during the second half of this century, discussions have been limited by the definitions used and the complexity of trust itself. The thesis gives a useful tool to those who work with trust, not because it is a *de facto* statement of what trust is, rather because it provides a tool for clear discussion.

The literature available to those researching trust is wide, yet on the whole is not mainstream. The thesis provides a summary of the trust literature.

5. Method

We have adopted a natural science approach to trust: it is essentially a natural, or given phenomenon, and as such can be subject to particular methods in its examination (Popper, 1967; Popper, 1969). Our approach has the following benefits:

- It circumscribes trust: the type of trust we address has clear boundaries, and the formalism makes them explicit.
- It is simple: we use Occam's Razor — we would prefer to be obviously wrong rather than imprecise.
- It (including the experiments discussed in chapter 7) can be easily replicated.
- It is not *final*: at some point the work here must be documented. Because of its nature, there will be questionable points within the thesis. This is a strength, since by questioning these points and suggesting alternative approaches, the formalism is a route to increasingly solid representations of trust.

The methodology used in this work is discussed in chapter 2.

Declaration

The work in this thesis is the original work of the author. The idea of a formalisation for trust came in the creative atmosphere of the Canon Research Centre in Guildford. Its development was spurred on by the staff there and in the Department at the University of Stirling. The formalism presented here is solely the author's.

Early realisations of the formalism have been published elsewhere, in Marsh (1992, 1993), and Marsh and Thimbleby (1992). A recent publication (Marsh 1994b) presents a revised version of the formalism, and in Marsh (1994a), the formalism is applied extensively to considerations of dispositions, as discussed in chapter 4.

In addition, work on extending the formalism's application to CSCW has been presented in Thimbleby *et al.* (1994) and Jones *et al.*, 1994, the contents of which were based on the author's work on the formalism. The texts of Marsh (1992, 1993, 1994a, 1994b) are presented in appendix C, and Marsh (1993, 1994a) are available as Technical Reports from the Department of Computing Science and Mathematics, Stirling.

Acknowledgements

In a piece of work that takes as long as this, there are always people to thank. In the past four years, there has always been someone prepared to help me. My apologies to those I may miss — a collective Thanks to All.

My supervisor, friend and mentor, Harold Thimbleby, has known when to be one and when to be another. His guidance and keen insight, along with his thirst for new knowledge and excitement when approached with new ideas have made this time fun and always stimulating.

This work was sponsored by Canon Research Europe under a SERC CASE award, to whom thanks are due. The time I spent with Canon saw the ideas in this work germinate, grow, and blossom due to their careful insights and keen intelligence. Many thanks in particular to David Lau-Kee for setting this up.

The Department at Stirling is made up of dedicated and insightful people. My thanks to Ian Wilson for asking awkward questions and more often than not supplying the answers. (The dichotomy of trust as black and white *vs.* trust as a measure is due to him.) Chic Rattray has been a source of useful information. Jane and Moira have laughed at my jokes, even at 9am on a Monday — thanks for dedication beyond the call of duty! To the other postgraduates in the department, thanks for asking me to lunch! Sam, Graham and Catherine, the computer officers here in the department, have been a source of information and the odd beer (no, not at work).

Lynne Coventry, Steve Jones, and Andy Cockburn, coffee bar buddies and office mates, latterly colleagues in 'other places,' have been a source of inspiration, beer and other alcoholic niceties, and friendship. Sharing an office with Andy and Steve is an experience not to be forgotten... To all my friends, both in and out of university, thanks for the phone calls, the tea bags, the beer, and the fun.

My family, Ann and Arthur (Mom and Dad), Jo and Peter, have always been there. Their love and support has been a foundation for my life. My parents-in-law, Margaret and Robert Young, who accepted me into their family without hesitation, and who have helped me in too many ways to mention since we met, deserve thanks and a medal for their support (and the odd bike lesson!).

The final word of thanks goes to Susan. Her love, understanding, moral support, care, and very existence are my keystone. I can honestly say that without her, I would not have finished this work. Most of all, thanks for changing my life on August 14th, 1993.

This work is dedicated to Susan Marsh. Here's to an everlasting trust between us.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> ... trust is a social good to be protected just as much as the air we breathe or the water we drink. When it is damaged, the community as a whole suffers; and when it is destroyed, societies falter and collapse.
>
> Bok, 1978, pp 26 and 27.

1.1 A Brief Preview

Suppose someone offers to help you fix your broken down car on the motorway. You've never met them before, but they're wearing garage overalls, and they turned up in a pick-up truck which you saw a few minutes ago at a service station. Do you accept their help? Unless you're a member of a motoring club, the answer is most likely to be yes. Now suppose the guy was in jeans and a T-shirt, and turned up in an old VW Beetle. You would probably take a lot more time in accepting help, asking lots of questions, such as who the man is, where he's from, and so on. In the first instance, you trust the mechanic because of what he is, or at least, what you can see he is. In the second, in order to have any serious trust in the man in jeans, much reassurance is necessary.

We all make trusting decisions, most of us every day of our lives, and many times per day (Luhmann, 1979). Each time we make such a decision, we put something on the line — our lives, our house, a book: something. The decision to trust is based on evidence to believe, or be confident in, someone or something's good intentions towards us (Yamamoto, 1990). Sometimes, the trustee (to widen the accepted usage of the word) lets us down, or betrays our trust. When that happens, they are trusted less, if at all. Betrayal of our trust is not our responsibility, rather it is that of the trustee (Hertzberg, 1988). This is because "trust can only concern that which one person can rightly demand of another." (Hertzberg, 1988, page 319).

This thesis is concerned with the introduction of a formalism for trust. It approaches the concept from the point of view of an artificial agent making its way in the world. Thus, it represents a departure from other work involving trust, whilst remaining applicable to that work. In other words, the formalism can be used as part of the overall structure of a rational, intelligent, trusting artificial agent, but it also provides those who study and discuss trust with a means of discussing it in a precise and straightforward manner. One area of interest is that of Distributed Artificial Intelligence (DAI), or Multi-Agent Systems (MAS),[1] the inherently social study of artificial agents existing with each other in some society, or 'world.' The word 'social' is important here — most of the trusting decisions we make in the world concern others, although occasionally they concern ourselves or the environment in which we exist (Luhmann, 1979). In that sense, trust is a social phenomenon, and is present wherever societies exist (Yamamoto, 1990; Baier, 1986). It is from that viewpoint that we approach DAI/MAS — since there must exist a society of agents (whether we like it or not (Gasser, 1991)), then trust, implicit or explicit, is present. At this juncture, the trust is implicitly assumed — "It is absolutely essential ... that the agents are known to be trustworthy; the model would have to be developed further to deal with shades of untrustworthy behaviour on the part of agents." (Rosenschein, 1985). In other words, the agents that are at present being designed *assume* trust; an assumption both unreasonable and misguided (Hertzberg, 1988). This thesis introduces an explicit trust to agents, to allow them to reason with and about trust, thereby making them more robust in the face of decision making concerning others.

The remainder of this chapter is concerned with introducing trust, its uses, some definitions, and a justification for the formalism. Cooperation is also discussed, and how it relates to coordination and collaboration. Aspects of trust are presented from several points of view, and a brief description of the remainder of the thesis is presented.

[1] For the remainder of the thesis, we treat the two terms as synonymous.

1.2 Unravelling DAI/MAS

There are aspects of DAI which need clarification, particularly with regard to how it touches on other fields, notably Artificial Life (AL), Computer-Supported Cooperative Work (CSCW), Sociology, Anthropology, and Social Psychology. The list is longer than this — DAI is a field of research perhaps unique in the number of aspects it bears relevance to. The thesis will thus be of interest to practitioners in many other areas of research, since it provides them with a general, abstract method for discussing and reasoning about trust — something which is applicable to social sciences of all types[2]). DAI is included as a social science because at its heart lies the need to understand why some things happen in societies, or groups, of agents, and to harness the power group action can give. For example, it is quite likely that two agents will clean up a room quicker than one (although not necessarily). This begs questions such as why and how, why with each other, and so forth. Such questions are inherently applicable to sociology and psychology, and also to DAI. Acknowledging the absence of answers, this thesis attempts to provide a means of attaining some; it also provides a means for reasoning about behaviour, and for deciding on behaviour for agents.

1.3 Cooperation, Coordination, or Collaboration?

In DAI, cooperation amongst agents is of interest (perhaps of ultimate interest), thus a key thrust of this work concerns cooperation. We consider trust to be a major aspect of cooperation of any kind. We make a distinction between cooperation, coordination, and collaboration, since the three are closely related. We take our definitions from Jones (1990), who states the following:

- **Cooperation** is when people (agents) "work and act together on a task, and share the profits or benefits from doing so." (page 3).
- **Collaboration** is where "one works jointly with other on a task." *ibid.*

[2]We define a social science to be a 'science' concerned with aspects of social existence. Clearly AL fits into this frame, despite its biological bent (see chapter 3). CSCW is a restricted social science in this sense, since it deals primarily with only one aspect of social behaviour, that of cooperation. Nevertheless, its focus is clearly social.

- **Coordination** is "the act of causing parts to function together or in a proper order." *ibid.*

From this point of view, cooperation and collaboration are very similar. They differ since in cooperation, individuals work more independently and exchange information (or help) where necessary (Jones, 1990). Coordination is essential in a cooperative or collaborative task, but we address it in a cursory manner only in this work. We are more concerned with aspects of cooperation, such as why it should happen, and to what extent trust plays a part.

1.4 Aspects of Trust

1.4.1 Defining Trust

Trust is a common phenomenon. Indeed, it has been argued that we as humans would not even be able to face the complexities of the world without resorting to trust, because it is with trust that we are able to reason sensibly about the possibilities of everyday life (Luhmann, 1979). For example, we leave the house every morning *trusting* that we will be able to return, and will not end up in hospital because of some accident that we *trust* will not happen. Despite its importance, there has been a lack of detailed research on the topic (Golembiewski & McConkie, 1975; Luhmann, 1979; Luhmann, 1990). In addition, the work that has been carried out presents its own problems, not least that a solid accepted definition of trust still eludes us. The definition given by Morton Deutsch in 1962 is more widely accepted than many, and states that trusting behaviour occurs when an individual perceives an ambiguous path, the result of which could be good or bad, and the occurrence of the good or bad result is contingent on the actions of another person; finally, the bad result is more harming than the good result is beneficial. If the individual chooses to go down that path, he can be said to have made a trusting choice, if not, he is distrustful (Deutsch, 1962). This definition is acceptable to the extent that the basic structure of a trusting choice is shown. There is disagreement as to the idea that the benefits should be less than the harm done, however: Golembiewski and McConkie (1975) present a similar definition, but stating that "the loss or pain attendant to unfulfillment of the trust is *sometimes* seen as greater than the reward or pleasure deriving from fulfilled trust." (*ibid.*, page 133, my emphasis). We proceed to accept the idea that:

> Trust implies some degree of uncertainty as to outcome.
>
> Trust implies hopefulness or optimism as to outcome.
>
> Golembiewski and McConkie, 1975, page 133.

Trust is thus strongly linked to confidence in some thing, be it the person to be trusted, the environment, or whatever it is that the desirable outcome is contingent upon.[3] We arrive at the concept of trust as choosing to put ourselves in another's hands, in that the behaviour of the other determines what we get out of a situation. This is further discussed in chapters 3 and 4. The following sections are concerned with presenting some aspects of trust, in terms of social, technological and biological considerations, before we proceed to discuss the area of DAI in more detail.

1.4.2 Social Aspects

In societies, trust is a fact of everyday life (Yamamoto, 1990; Baier, 1986; Deutsch, 1973; Luhmann, 1979; Luhmann, 1990). Indeed, without trust, as the opening quote of this chapter suggests, societies would cease to exist (see also Lagenspetz, 1990). There are many examples of where trust plays an explicit role in societies. That we get up at all in the morning is a sign of the trust we have in society and our environment (Luhmann, 1979). Since this thesis is concerned with agents working together in some way, we touch on two of the stronger examples. These are firstly groups, and how they are formed and exist, and secondly the phenomenon of working together with others. This could be in a simple fashion, such as moving furniture, or in a way which involves some extremely complex relationships, for example in a hierarchical organisation, such as a large business.

Groups

Group formation is in itself a complex area, one which it is beyond the scope of this thesis to address in a detailed fashion. This section suggests some aspects of groups which trust may relate to. Much more needs to be done in this area (see chapter 8) but this section represents an introduction. Its form is that of a series of questions which trust can help in providing answers to. The questions are as follows:

[3] There are more forms of trust than trust-as-confidence. These are discussed in more detail in chapter 3.

1. How much does trust come into the process of an individual joining a group, both from the point of view of the group, and the individual?

   In other words, how much does trust matter for the group to allow an individual to join, and how much does trust matter for the individual to want to join the group in the first place? Answers to these questions are already partly available. Individuals generally join groups to further some particular goal (Shaw, 1981). The aspect of trust in this is uncertain, other than the individuals 'trusting' the group to be able to achieve its goal(s). From the point of view of the group trusting the individual, easy answers are not available.

2. How much does trust help in determining intergroup behaviour?

   Intergroup behaviour is a difficult and complex subject. Groups have been found to behave in an aggressive manner towards members of other groups (and thus the groups themselves) with which they are in competition (Brown, 1988). Since we are primarily concerned with cooperation, it is educational to consider intergroup cooperation. According to Brown (1988), intergroup cooperation increases positive intergroup feelings as long as the outcome is success, but the story is different if the outcome is less than successful. Where trust may come in is as yet an open question, unless we perceive groups as individual entities. In that case, we could ask whether trust increases on positive encounters, and decreases on negative ones, or if there may be more depth there.

3. How much of a determinant of intragroup behaviour is trust?

   With groups, how do members react to one another, and how much does trust play a part in this? Do members trust each other more than non-members? If so, why and to what extent? All of these remain somewhat open questions, but all involve trust, and would benefit from a more precise method of discussing and representing the phenomenon.

The list of questions is not exhaustive. It is sufficient to see that trust may well play a part in group formation, behaviour, and structure. In addition, because the formalism introduced here is intended for embedding within artificial agents, we can observe group behaviour amongst such agents, and the behaviour of the trust between them.

Working Together

To a large extent, working together is closely related to working as a group. We separate the two here because there are aspects which can be considered differently. For example, why one agent, or person, decides to work with another, and how much he needs to trust her in order to do so. Since working together invariably implies one of the 'C' words — Cooperation, Coordination, Collaboration, and generally Communication, it is of interest, and is discussed with the use of the formalism in chapter 5.

1.4.3 Biological Aspects

Is trusting behaviour a uniquely human phenomenon, or do other members of the biological world exhibit trust? The answer to that lies largely in delayed reciprocation in the animal world (Harcourt, 1991). Animals help each other out in the hope that, in the future when they need help, they will be helped back. A good example is the vampire bat. When such bats have had a good night, and a surplus of blood, they feed those who have not (Harcourt, 1991). Invariably, those who behave in this way are fed by those they feed when they themselves have a bad night's hunting. Is this an example of trust? Clearly, the bat who feeds others first is displaying some primitive form of trust in those he feeds, because he is trusting them to reciprocate in the future. Because time matters in the real world, and things take place over time, there is an element of risk in such behaviour. "This incorporation of risk into the decision can be treated under a general heading that can be described by the single word 'trust.' Situations involving trust constitute a subclass of those involving risk. They are situations in which the risk one takes depends on the performance of another actor." (Coleman, 1990, page 91). Thus, the bat displays trust. Of course, the bats who continually reciprocate as a group attain an evolutionary advantage over their non-reciprocating neighbours, since those who reciprocate will be fed when they need feeding, thus increasing their (and their reciprocating friends') chances of survival in the long term (Harcourt, 1991).

There are other examples of delayed reciprocation, such as chimps helping each other in fights (Trivers, 1985). Indeed, Harcourt argues that trust is more likely to be present here since the intellectual capacity of primates is greater than many other animals.

Much of the reasoning behind delayed reciprocation can be addressed in terms of 'survival of the fittest genes.' Animals help those who are related to themselves because it perpetuates their genes' chances of survival, or *vice versa*, in that those who help others already do have genes which are predisposed to helping others (for more discussion on this extremely interesting topic, see Dawkins (1986, 1989, 1990)).

1.4.4 Technological (Artificial) Aspects

In the modern world, computers are becoming all-pervading. An argument of the present work is that where a society exists, so does trust. Artificial societies already exist, in however limited a fashion. An example of such an artificial society is the phone network. It consists of many nodes, each of some intelligence, each deciding which way to route traffic (phone calls, faxes, etc.). The very fact that these nodes work with each other means that trust is present. It is, perhaps, a limited trust — 'should I route this message via that node, or not' could translate to 'do I trust this node enough to route this message through it?' Another example is the Internet. Just how much trust comes into such systems was evident in the virtual collapse of the Internet a few years ago (Spafford, 1989). That it collapsed was perhaps due to excessive laxity on the part of some people, but the Internet Worm took advantage of various 'trusted host' connections (Spafford, 1989).

Trust is something little understood in such artificial networks (Woo & Lam, 1992). The formalism presented in this thesis would allow nodes in such networks to reason with and about trust, but also would allow network managers another means of assessing their networks. Of course, such a means of assessment would need careful handling — explanations, much as in present expert systems, would have to be provided to show how decisions to trust or not to trust were arrived at. Whilst the formalism is capable of handling such requirements, they are outwith the scope of this thesis.

DAI/MAS

DAI is social (Gasser, 1991). This sociality means that trust is involved at many levels. Since the agents in DAI are much more autonomous and independent than those in, say, a phone network, the considerations they must make regarding others are similarly more complex. When the others are human (see Thimbleby *et al.*, 1994, Jones *et al.*,

1994) the considerations become ever more complex. At present, there is no explicit consideration of trust in DAI. We argue that this is short-sightedness on the part of DAI. If we are to expect our agents to survive in the 'real world' (i.e., outside the sterile confines of the research laboratory), we must make them more robust in respect of their interdependence with others, their reliance on others. Relying on others' good behaviour is not enough (Marsh, 1992). An argument of the present work is that the incorporation of trust into an intelligent agent's considerations of others will be a step forward in providing the needed robustness, without losing any of the freedom of choice that such agents will be expected to possess.

1.5 Trust — Justifications and Arguments

The phenomenon of trust is difficult to 'pin down.' At the heart of the problem lies the fact that all of us, as human beings, trust. Thus we all have an idea of what trust is (Golembiewski & McConkie, 1975). A fundamental justification for the thesis is that being able to discuss trust in a precise manner would be an important step forward in our understanding of the phenomenon. In other words, being able to argue about our view of trust in language which everyone else can comprehend and discuss should help us reach a solid definition of the concept. Failing that, we may find that trust is simply too complex to formalise in the fashion presented in this thesis — there are too many features which are not generalisable, or require more complex formulæ. That said, this would still be of importance since without the attempt, answers to questions such as whether trust can be formalised, or whether it is a notion which is too based in the emotions, would not be forthcoming (see the following chapter for a discussion of this).

A further justification relates to DAI itself. As was discussed above, DAI assumes trustworthy behaviour at present (Rosenschein, 1985; von Martial, 1992). This works well in a laboratory, since we mostly study and expect cooperative behaviour. It will not work in the 'real world,' since not everyone is trustworthy, and they may be malicious. An understanding of trust for an agent would provide two major benefits:

- It would allow the agent to consider others with respect to trust, enabling sensible, informed decisions to be made about who is or is not trustworthy.
- It would allow an agent to consider aspects of a particular situation, notably with

regard to who to accept help from, or ask for help from, and who to cooperate with if necessary.

In order to allow such reasoning, we have to embed the concept of trust into an agent. A first essential step in this direction is the development of a formalism and associated formulæ which are executable by an agent, i.e., able to be embedded within that agent. The primary aim of this thesis is the development of such a formalism. The formalism, though, is just a means to an end; some end results, in the form of trusting agents, are presented in chapter 7.

1.6 Overview of the Thesis

This thesis addresses some of the problems and solutions outlined above. The following chapter presents a deeper discussion of the methodology used here. In addition, the decision to use quantitative instead of qualitative data for the values in the formulæ is critically discussed. **Chapter 3** presents a comprehensive survey of much of the literature to be found on trust. The end result is a deeper justification of the requirement for the consideration of trust than is presented above, whilst providing an understanding of what work has been done with trust.

The formalism is a major contribution of this thesis. It provides the social sciences with a valuable tool for the precise discussion of trust, whilst it also gives fields such as DAI the means to embed trust within agents. The formalism is presented in **chapter 4**. This chapter introduces the main concepts of the formalism and associated formulæ, and goes on to extend the formalism to take ideas such as reciprocation into account. Having been introduced, the formalism is used in **chapter 5**. Examples of the formalism at work are presented, together with discussions about some of the difficulties which may arise. **Chapter 6** presents some rules and principles which trust appears to obey, in terms of the formalism and in an informal manner. **Chapter 7** presents a discussion of some of the results obtained from implementations of the formalism.

The need for further work and exploration is necessary in any useful research. This work, however, is different in the sense that it opens many doors onto possible work. The chapter on further work (**chapter 8**) is extensive, providing detailed discussions of work that can be done, and different ways of looking at trust. It is hoped that the work presented here will act as a spur for more research to be carried out in the area. **Chapter 9** summarises the results and discussions presented in this work.

1.7 Summary

- This thesis asserts that the incorporation of trusting considerations in artificial agents will make them more robust in the face of social uncertainty. For example, although the future is uncertain, decisions have to be made, often in the light of incomplete and possibly incorrect information. A knowledge of the workings of trust, and a trust in others, the information the agents receive, or in their beliefs, would allow them to make decisions which are either informed or acknowledged to be risky, along with measures of the risks involved.
- The development of a formalism which is straightforward and precise will allow such an incorporation of trust into agents.
- The formalism has several other benefits:
  1. It provides a tool for the discussion and clarification of trust.
  2. It is in itself discussable. That is, using the formalism, we are able to ascertain whether or not what we are representing *is* 'trust,' and modify the formalism accordingly.
  3. It may prove useful to other areas where sociality is prevalent, for example distributed systems.

Chapter 2

Methodology and Related Aspects

2.1 Discussion

With a subject such as trust, it is difficult to determine that what is presented is what is professed to be presented. There are many reasons for this, some of a subjective nature, some philosophical, some scientific. The way the subject has been approached in the remainder of this work inevitably suffers from many of these problems. However, other approaches would suffer also, and perhaps not so clearly, to their detriment. This chapter discusses the methodology used for the work carried out here, and assesses its strong and its weak points. Additionally it presents a reasoned discussion which, whilst taking into account the arguments against what has been done in this work, suggests that the formalism presented in particular is of use to social scientists and DAI researchers alike. In addition, it is useful because, should it fail, it presents a 'start' to proceed with, or a 'finish' with which to accept the elusive, uncapturable nature of trust.

2.2 The Methodology, and a Critique

2.2.1 On scientific contributions, according to Karl R. Popper

Trust is a natural, or 'given' phenomenon. As such, it has been approached here from a point of view in accordance with that of Karl Popper (Popper, 1967; Popper, 1969). In his work, Popper asserts that a scientific theory has a certain structure (Popper, 1969). A scientific theory, much as the formalism in this thesis,[1] must satisfy the following criteria:

- It must *demarcate* the area from pseudo-science. In other words, the theory must be *testable*, *refutable*, and *falsifiable* (Popper, 1969, pp. 36–39). Each proper test of the theory is an attempt to falsify it. The theory of trust presented in this work is inherently testable, since it stands as a formalism. What the formalism does in addition is to circumscribe the area to which it applies: it is clear that the formalism applies in certain circumstances, or types of trust, and not in others. The formalism structure *itself* makes this clear. It is only within the boundaries to which it applies that it can be tested (Popper, 1967).

- The theory must be *simple*. This is an application of Occam's Razor. Indeed, Popper states that this is what he calls 'Berkeley's Razor,' and states that "This razor is sharper than Ockham's." (Popper, 1969, Page 171). Simplicity is better than complexity for many reasons: firstly, it allows extreme tests to be carried out on the theory, which admits it to being more scientific than more complex theories (Popper, 1967; Popper, 1969). Secondly, without attempting a simple exploration first, little can be said about the results obtained: without doubt, a complex formalism, based on advanced concepts in mathematics, could be applied to trust, but there are problems here:

  - If it works, we can never *know why*: it may work because that is the only way of formalising trust; it may work because of an emergent or chaotic property of the formalism itself which cannot be readily explained.

  - If it does not work, there is little that can be done to *explain why*, since any part of the complex formalisation may be at fault.

  The simple formalisation allows for strict testing, and if it <u>should not work</u>, points to the need for a more complex formalism, or to the fact that the theory (that trust can be formalised) is incorrect. If it <u>should work</u>, we can be sure that there are no hidden reasons why: "the simpler theory has always a higher degree of testability than the more complicated one." (Popper, 1969, page 61).

[1]The formalism presented here is a step in the direction of increasing knowledge about trust. As such, it presents itself as a *theory* about trust, its workings, and its behaviour.

- It must be capable of replication, or duplication. Results, when given, must be able to be repeated, with the same results found. Without this, the theory is not scientific: "Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence,' but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable" (Popper, 1967, page 45). The formalism given in this thesis is both testable and reproducible: results from the experiments carried out using the formalism can be easily and quickly reproduced.

Finally, scientific theories are never static, they are "perpetually changing. This is not due to mere chance, but might well be expected" (Popper, 1967, page 71). In fact, "The game of science is, in principle, without end. He who decides one day that scientific statements do not call for further test, and that they can be regarded as finally verified, retires from the game." (Popper, 1967, page 53).

The formalism presented here is in flux, and can be continually adjusted and refined. It may be possible to question a formula on, say, page 117 at some point in time. This is not a fault of the formalism, rather it is one of its strengths as a scientific theory. Questioning and altering such a formula will refine and strengthen the application of the formalism. Therefore, a distinction must be made between the detailed *contents* of the thesis and its *contribution* to understanding trust — it is only with the presence of the formalism that we can question single formulæ, and were it not presented, we would be no further towards understanding trust.

The presentation of trust in this thesis is therefore a risky undertaking: the formalism can not be finalised, and thus can be explicitly questioned. This is also a strength. The taking of risk is *necessary* in science (Popper, 1967). The formalism may be partially or completely incorrect, but its presentation has furthered the understanding of trust.

The contribution of the formalism as a theory of trust is thus:

- Circumscription.
- Simplicity — Occam's Razor.
- Replicability.
- Flexibility, or 'non-finality.'

2.2.2 Attacking the Phenomenon

The previous chapter raised some definitions for trust, arriving at the idea that trust involves putting oneself in another's hands, with some hope as to the beneficial outcome of a situation. There are those who would disagree with such a definition. For Deutsch, the definition would be too weak, since it does not address the idea (for him) that the costs must outweigh the benefits (Deutsch, 1962). Others would think it too strong, since putting one's self in another's hands is a severe notion, and, in any case, we may have *some* form of control over the actions of the other. Both arguments are valid, both miss the point in a subtle manner. To some extent, the approach taken here does not need a *de facto* notion of what trust is. This is because of the path we take in defining the formalism. With a phenomenon as difficult to grasp as is trust, which is evident from the number of definitions we have of it (see the previous chapter, also chapter 3), it is of use to take an 'end-state' view of the concept. In other words, we observe the phenomenon, and attempt to define a formalism which behaves in the same manner, any definitions which have been put forward notwithstanding. The definitions are of use because we can test the formalism using them, or *vice versa*. Were we to choose a particular definition, basing the formalism on that, it would suffer from the inadequacies of that particular definition. As it is, since it does not depend on any one definition, there are two possibilities: we can derive all of the benefits of many definitions, or we can suffer from all the different drawbacks of each. Either way, we can test the results against each definition, and against expected trusting behaviour, from a subjective or an objective point of view. The methodology is of interest because of this.

2.2.3 Discussing the Methodology

One of the problems inherent in discussing trust is that the phenomenon is such a subjective one (Golembiewski & McConkie, 1975). It is difficult to grasp even our own private notions of what trust is, even without arriving at a general definition. This is where many of the previous studies have suffered. In attempting to provide a detailed definition, they have restricted themselves to just one or two of the several aspects trust shows, although some of them acknowledge this (Deutsch, 1962; Golembiewski & McConkie, 1975). From the point of view of studying the phenomenon in limited terms, this is not a major problem. Where the present study differs is that, instead of taking a particular definition, we adopt an approach which studies the essential behaviour of trusting, and attempts to capture some of the essence of that in a formalism. So, instead of starting with a definition and heading for a formalism, we start with what are by and large intuitive ideas about how trust works, based on experience[2] and the literature available, coupled with many working definitions of trust, and attempting to develop a formalism around them. This approach has its own drawbacks — not least, the ideas with which we develop the formalism may well be wildly wrong. In addition, the formalism may suffer from being too biased towards one of the aspects of trust, not touching another. The advantage we have is that we may at least know of these biases and drawbacks, and can acknowledge their existence. This is of great utility when refining or discussing the workings of the formalism.

Assuming that a formalism can be and has been developed using the above strategy (which is essentially a bottom-up approach compared to the top-down approaches of definition to discussion) it is simple enough to ascertain the use of such a formalism. For any definition of trust, then, it is possible to devise a test of the formalism to see if it fits in with the accepted definition. One of the strengths of this approach is that this stage could be seen as a refining stage, gradually bringing the formalism closer to an agreed, or accepted, point at which it does represent most people's views of trust — what it is, and how it works.

To illustrate this point, we digress.

There is, in the sphere of Artificial Intelligence, an ongoing argument about what AI actually *is*. See for example Schank (1991) and the Communications of the ACM, March 1994. The problems are twofold. The simplest of the problems, but perhaps the hardest to address, is that the goalposts are continually changing (Schank, 1991). Not so long ago, a world-class chess-playing program would have been seen to be AI. Not any more. This is to be expected — after all, people's perceptions change as the examples with which they are presented become more detailed and complex. The second problem is of more relevance to this discussion. It concerns what intelligence actually is: how it works and how to represent it. There are, within AI, several schools of thought, from the view that intelligence is the manipulation of symbols (Newell & Simon, 1976) to the ability to pass the Turing Test (Turing, 1950).

Needless to say, AI has problems regarding its identity (Schank, 1991).

[2]Another of Popper's criterion for good science (Popper, 1967).

The same conceptual problems apply to the trust formalism presented here. "What is trust?" is a question which is particularly difficult to answer, and it is sensible not to try to answer it in any fashion which is not empirical. Thus, the formalism is developed from experience, intuitive expectations about trust, from a subjective point of view, and conclusions to be found in the sociological, psychological and philosophical literature which relates to the subject. In this way we amalgamate definitions, intuition and observation, avoiding the problems of particular definitions, and simply create a formalism which should behave the same way trust does. In other words, we make no claims about the validity of the workings of the formalism, although they are based along the lines mentioned above (the philosophical literature has had a deep impact on the formalism). What we do claim is that the end result of the application of the formalism to a consideration of trust is 'as if' we had been considering the problem using 'real' trust.

There are problems with this approach. We stand to suffer from the same kind of problem that any more formal definition suffers from, namely that the kind of trust we define (implicitly or explicitly) using the formalism, that is, the kind of behaviour we produce from trusting agents, is indicative of only one sub-class of trust (and there are many — see the following chapter). Indeed, that much is to be expected, since, however abstract the formalism may be, it still has to be founded on some aspect of trust, which is one of many types. However, the resultant trusting behaviour is sufficiently 'like' trust that it is applicable in several domains, thus representing several of the sub-classes of trust. In addition, the fact that the formalism is there at all allows us to alter, correct, and improve it to take in more definitions and sub-classes of trusting behaviour.

2.2.4 What if the formalism doesn't work?

Thus far, we have assumed that the formalism developed here is correct. Underlying this assumption is a deeper one, that trust itself is capable of being formalised. In other words, we hypothesise that it is possible to isolate 'trust,' capturing its essence, and putting the results into a formalism. This is a major step forward in the understanding of trust. So, what happens if this can't be done? Trust itself may be so closely linked with other aspects of the human psyche, such as morality and justice, that isolating it is impossible.

One of the reasons for working the way we have is that it is possible to come up with results fairly quickly. Since we are attempting to model end results, we need have no precise notion of the interim phases of the concept. That being the case, it is a relatively simpler affair to formalise trust (an application of Occam's Razor). The resultant behaviour is, we suggest, representative of trust, but not, we accept, reflecting a deep consideration of the aspects of a particular situation.[3] Deeper consideration will show the utility of allowing experience of past situations and behaviour to guide a trusting decision (as with 'real' trust) and the benefits of such an approach in terms of speed, efficient consideration of possible outcomes, and preciseness of decision.[4]

We return to the point of the formalism not working at all, or failing to capture crucial aspects of trust. There are several considerations still to be made. As was mentioned in the previous chapter, the end result of such a piece of work may be that we have to conclude that trust is simply not formalisable, or that a more complex formalisation is necessary. Although a disappointing result, it is no less important. Such a result would be useful for many reasons, not least because it would allow time to be devoted elsewhere. Also, it may open avenues for research into other human traits, such as emotions. Nevertheless, it is difficult to prove such a statement as 'trust cannot be formalised' without having tried in the first place. Indeed, such a statement may be self-contradictory since, in order to put forward such a hypothesis, a formalism must have been presented which represented trust well enough for us to be able to say that it was not good enough, and it must also be tested within its own demarcation lines (Popper, 1967). What may be able to be said is that the way in which we have approached the problem does not, and will not, bear fruit.

[3] Here lies another of the formalism's strengths: if we were to consider trust in particular situations, we would most likely be limiting ourselves to those situations. For the formalism to be of any use, it must be generalisable. That this results in a loss of specificity is, from this viewpoint, no bad thing.

[4] The final aspect needs some clarification. Since we elected to use values for our trusting formalism, we eventually come out with one number with which to make our decisions regarding trust. This is discussed further below, but in terms of being precise, this number (value) for trust is inescapably that. Indeed, it is perhaps over-precise, but see further below.

2.3 Discussing the Use of Values

In this work, we have chosen to represent trust as a continuous variable over a specific range (here, $[-1, +1)$). We discuss the benefits and drawbacks of such an approach here.

2.3.1 Sensitivity

In a formalism using quantitative data, it is possible that small differences in individual values produce relatively large differences in the overall result (cooperate or not, trust highly or not, and so forth). This could be perceived as sensible behaviour in trust — certain basic values, such as general trust in agents, or the risk perceived in a situation, are of course important factors to be taken into account, and small variations in these should, it could be argued, produce large differences in final results. That these basic values are somewhat more inflexible (as in basic trust (Baier, 1986; Golembiewski & McConkie, 1975)), or perhaps more important (as in the risk inherent in a situation) reflects the idea that small differences in them should alter the result by a larger amount. We consider this further in chapter 5, once the formalism has been introduced and some worked examples given.

2.3.2 Subjectivity

Agent-subjectivity in the use of values is a major problem. It is possible to imagine, for example, one person placing a value of 50% on how much he trusts another. So far, we have a straightforward value placed on trust. The problem arises when that truster tells another how much he trusts the trustee. The third man takes the 50% value on hand, but here he sees 50% as very high, trusting the trustee with more, perhaps, than the first would deem sensible, since for the first 50% is merely an average trust value. To counter the argument against subjectivity in values, it is possible to use a stratification of trust, a kind of fuzzy logic of trust, giving each strata a label, so for example a trust of (for argument's sake) $+1$ would be labelled 'blind trust.' A suggested stratification is given in table 2.1.

Table 2.1 is a simple stratification, and bears a similarity to fuzzy logic (Kosko & Isaka, 1993). The advantages of this stratification are that a trust designated as 'high' by one agent is acknowledged by others as a high trust. Thus we avoid the problem of

Value range	Label
$+1$	Blind Trust
> 0.9	Very high trust
0.75 to 0.9	High trust
0.5 to 0.75	High medium trust
0.25 to 0.5	Low medium trust
0 to 0.25	Low trust
-0.25 to 0	Low distrust
-0.5 to -0.25	Low medium distrust
-0.75 to -0.5	High medium distrust
-0.9 to -0.75	High distrust
< -0.9	Very high distrust
-1	Complete distrust

Table 2.1: Possible stratification of trust values

'what does a trust of 0.5, or 50% mean? Is it high or low,' for example. We still suffer from agent-subjectivity, however. Naturally, different agents will see different things in specific labels. So, an agent who assigns the value 'high trust' to the trust she has in another will pass that value on to another who, perceiving it as high, actually sees this 'high trust' as what he would ordinarily say was 'medium trust', since the metrics each uses may be different. These are problems which will endure, however. A more difficult problem is the loss of sensitivity and accuracy when using strata. Once an agent is judged to be trusted 'highly,' then another agent who is more trustworthy, but not enough to be in another strata, suffers in the comparison. In addition, when we reach a value just below blind trust, call it 'extremely high,' then there can be no-one trusted better, thus we suffer from the same problem as with blind trust, a value of $+1$ (see chapter 4). With values, there is always a higher trust that can be reached, as long as it is an open interval.[5] This acknowledges the variable and continuous nature of trust which strata cannot represent. With strata, unlike continuous values, it is not possible to say 'I trust him more than her, by a small amount,' because there will always be smaller or larger amounts with values, not with strata.

[5]This follows from the definition of real numbers.

Benefits	Drawbacks
Reflects continuous nature of trust	Problems with subjectivity lead to misunderstandings
Allows easy implementation	More difficult to understand
Experimentation using differing formalisms	Sensitivity may be a problem
Ability to spot subtle anomalies	Fudging remains an issue

Table 2.2: Benefits and drawbacks of using values for Trust

2.3.3 Anomalies

When we discuss trust in terms of the formalism provided in this work, there is a possibility that we may be able to observe anomalies in trusting behaviour which, until now, have not been observed or noted. That this is a possibility rests on the fact that we have a formalism which enables implementations of trusting agents to be developed, thus allowing experimentation (and possible refutation (Popper, 1967)). We can then seek norms of trusting behaviour, how successful they appear to be, their relationships with other forms of behaviour, and so forth. The results of some simple experiments will be presented in chapter 7. Whilst the methodology adhered to in this work has resulted in a relatively limited set of implementations, due to the need to develop and correct the formalism before such implementation, it is presented only as a first step along the road to fully implemented trusting agents. The implementations described in chapter 7 act as a pointer towards what may be possible.

We have mentioned many of the benefits and drawbacks that may be gained from using values for the formalism. The main points are summarised in table 2.2.

2.4 Linearity

The formalism uses linear equations with which to estimate values for trust. One of the main considerations of the development of a formalism for inclusion in trusting agents was that the inclusion be a simple affair, uncluttered by problems of how to implement certain formulæ.[6] The decision to use linear equations was partly based on the need for simplicity and practicability in representing and modelling trusting behaviour on the part of the individual. There are problems with this approach however, one of which is a lack of flexibility. The previous sections discussed the event of trust not being formalisable as is; in other words, that the method we are using is at least insufficient for formalising trust. Of course, this does not preclude another formalism from capturing the essence of the phenomenon. Any other formalism may well be considerably more mathematically sophisticated than that proposed here. If that is the case, we lose simplicity but gain power, a common tradeoff. The use of linear equations is in the first instance justified by their greater practicability, extensibility, and predictability.

Another aspect of the linear equation is its ease of correction. With simple linear equations, mistakes are easily spotted, corrections quickly made, and the resultant formulæ open to all, and easy enough to understand. Since one aspect of the work presented here is its applicability to the social sciences as a whole, the wider the audience that can easily grasp the concept the better. Whilst complex equations look impressive, they do little for readability, and some simple mistake can easily be missed.

2.5 Conclusion

There is debate on the methodology used in this work, which has been discussed here. It is important to note that the methodology involves not only experiential views of how trust works, but also relies on the amalgamation of the great deal of work that has already been carried out with regard to trust. None of this work has been within the sphere of AI — trust has hitherto been seen as a topic for social psychology, for example. This work changes that view. A possible problem with the thesis is that its application to computer science may seem trivial. This is a result of the deliberate approach we have taken: we concentrate on simplicity, since the operational aspects of the problem are important. Computing Science does not have to be 'hard' to be applicable. Indeed, the simpler a system is, the more able we are to understand it, test it, and use it properly. The formalisation presented here has these benefits.

[6] Another application of Occam's Razor. This time applied to the embedding of the formalism in agents, rather than its conception.

The next chapter presents a detailed view of the work that has been carried out to date in social psychology, sociology and philosophy, together with other, more incidental treatments of the topic by researchers in other fields.

Chapter 3

Trust

> "Perhaps there is no single variable which so thoroughly influences interpersonal and group behaviour as does trust ..."
>
> Golembiewski and McConkie, 1975, page 131.

> "We inhabit a climate of trust as we inhabit an atmosphere and notice it as we notice air, only when it becomes scarce or polluted."
>
> Baier, 1986, page 235.

> "Trust ... is a basic fact of human life."
>
> Luhmann, 1979, page 4.

3.1 Starting Points

Trust is undoubtedly an important feature of our everyday lives. Without a background of trust, it has been suggested, we would suffer from a loss of efficiency and dynamism (Golembiewski & McConkie, 1975). Perhaps worse, we would find it very difficult to get up in the morning (Luhmann, 1979), and would suffer the inevitable collapse of our society (Bok, 1978; Lagenspetz, 1992).

So, what if trust *is* present? How do we benefit? We experience far better accomplishments in task performance (Golembiewski & McConkie, 1975), greater and more healthy personal development (*ibid.*), an understanding, or at least an acceptance, of the complexity of our society (Luhmann, 1979), and the ability to cooperate (Argyle, 1991; Deutsch, 1962),[1] to quote some authorities.

---

[1] Arguably, cooperation is viable when trust is not present, although the presence of trust, albeit in a limited sense, is a major promoter of cooperation among friends and strangers. See Argyle, 1991, page 34, for example.

There are many views of trust (Barber, 1983; Shapiro, 1987), and there are more than a few reasons for this. Two of these reasons stand out to be presented in particular, since they touch on the main problems that stand in the way of formalising the concept. Firstly, we are all 'experts' on trust, at least our own brand of it, and there is the problem, since, as there are so many different 'experts,' each of which could define trust differently, there are as many differing definitions, and thus views, of trust. This does not make life any easier when we wish to study the phenomenon. The second reason is more straightforward — there are as many views of trust as there are because there are a great many types of trust (Shapiro, 1987; Deutsch, 1973). Put another way, trust can be classified as hope, despair, confidence, innocence, and impulsiveness, to name but some (see Golembiewski and McConkie (1975), and Deutsch (1973)). Of these, only some are 'positive,' in terms of optimistic views of what is likely to occur. Trust in situations of despair, for example, is simply a reflection of the dilemma that the alternatives to trusting are so bad that trusting is the lesser of two evils (Golembiewski & McConkie, 1975). It is still, however, trust, although it is outwith the scope of this thesis.

As many types and views of trust as there are, there are also many fields which study the phenomenon. Thus, we find it mentioned in fields as diverse (it seems at first glance) as evolutionary biology (Bateson, 1990), sociology (Luhmann, 1979; Luhmann, 1990), social psychology (Deutsch, 1962), economics (Hart *et al.*, 1990; Dasgupta, 1990), history (Gambetta, 1990b; Pagden, 1990), and philosophy (Lagenspetz, 1992; Hertzberg, 1988; Wittgenstein, 1977). The question arises: are these examples as diverse as they may seem? Clearly, the view of the all-pervading nature and importance of trust can only be strengthened by such diversity of study. There is, however, a link that all of these examples discuss trust as part of a *society*, either a particular society (e.g., 18th century Naples (Pagden, 1990), or Accra, in Ghana (Hart *et al.*, 1990)) or society in general (e.g., Luhmann (1979, 1990), Baier (1986) and many others). The diversity of the fields of study is justified by their common factors, in particular the aspect of society.

Society, or the presence of some form of it, appears to be a salient factor in the appearance of trust. This, however, is a circular argument in a sense, since it has been argued that society depends on trust being present (Bok, 1978; Baier, 1986), see Yamamoto (1990) for a specific example of this. This certainly begs the question of which came first! Intuitively, it can be suggested that trust was there before society, at least in terms of what we perceive established society to mean. The definition of society is another contentious area (Frisby & Sayer, 1986), and one without a concrete answer. One definition comes from van den Berghe, who states that "society is a group of conspecifics bounded by a zone of much less frequent interactions than the rate which prevails between its members" (van den Berghe, 1980, page 77). Emile Durkheim saw society as a *sui generis* object, one which is irreducible to its members, and purely a moral order (Frisby & Sayer, 1986, especially chapter 2). This is akin to the properties of emergence in distributed systems (Forrest, 1990; Wavish, 1991), thus society is seen as emergent by Durkheim. Other sociologists, however, see it from a different angle, as "the network of shared understandings, the cognitive and communicative community which makes the actions of individuals ... meaningful to themselves and others" (Frisby & Sayer, 1986, page 75).

Definitions of society notwithstanding, it is clear that societies probably came into existence before trust. For example, although some degree of trust exists in animals, as shown by reciprocal altruism (Trivers, 1985, especially chapter 5. See also Trivers, 1971), or delayed reciprocation (Harcourt, 1991), we see that this appears in already established societies. To form a society, perhaps trust is not *specifically* necessary. For example, animals in need of protection through numbers form a society within which trust becomes important.

There exists, then, an ambiguity, brought on by definitions of society, of trust, and of experience and observation of trust, and its fostering or destruction in society.[2] Indeed, Ernest Gellner, in *Trust, Cohesion, and the Social Order*, states that discussions of trust:

> ... fluctuate between the notion of trust as something specific within a society, one thing among others, and another, broader version of it, which

[2]There is much to say here: On the definitions of society, see van den Berghe (1980), see also Campbell (1981) for a detailed review of several views of society. The observations of trust range through many of the sources already cited; see Gambetta (1990c) for a wide-ranging collection of essays on trust. Within that volume, there are discussions on the systematic destruction of trust, see Pagden (1990) in particular. See also Lagenspetz (1992), particularly the example in the first few sentences. The fostering of trust in society is of major importance in promoting peace therein, and work on (both inter and intra) group trust has been done — see Golembiewski and McConkie (1975), also Sato (1988).

> makes it coextensive with the very existence of the social order ... Trust as coextensive with *any* kind of social order is one thing, and trust as something within society, of which sometimes there is more and sometimes there is less, is another ... perhaps one way of making progress on the wider topic of social order as such, and the diversity of social orders, is to distinguish various kinds of trust.
>
> (Gellner, 1990, page 142).

There is also a dichotomy between trust as something which variates, and trust as something which simply *is*.[3] In other words, we often say 'I trust you,' but what does that mean? Does it mean, for example, that I trust you more than 60% of what I consider to be complete trust, or some other arbitrary figure which means that my trust in you is greater than some threshold value? (Below which, whether I have any trust in you or not, I do not trust you). Or is it a general statement of fact, which requires analysis on any action that should be taken. For example, I trust you generally, but that says little about *how much* I trust you in specific situations, like driving my car, or borrowing some money. It is easy, then, to say 'I *do* trust you, but...'

3.1.1 Overview of the Chapter

The opening section reviewed questions which are important, not only because trust is important in itself, but also because they raise fundamental questions about the *way* we trust. One thing that may have emerged from this discussion is that answers are few, far between, and vague at times. It is, as Gellner suggests, important to know what you are talking about before you start. Vague notions of trust are all very well, but a concrete basis for discussion would be of more help.

To that end, this chapter will cast some light on what others' views of trust are, much of the work that has been done, directly or indirectly, pertaining to trust, and what results and observations can be gleaned from them. The amount of literature on trust is substantial (Golembiewski & McConkie, 1975), although not necessarily mainstream (Luhmann, 1990). It would be an interminable task to sift through it

[3] I am grateful to Ian Wilson for pointing this out.

and present it in a meaningful form.[4] Instead, with the aim of providing a worthwhile overview of the subject, which does, at times, delve deeper into specific aspects of trust, we present the subject of study as one which naturally spreads itself across research boundaries, but has specific similarities within such boundaries. There is, however, one caveat: the work presented here is not intended to be a *de facto* statement of the research on trust; indeed, much work has been done, and is in progress, in many spheres, and much of this is presented below. The following sections depict the thinking of those whose work has had an effect, mostly direct, on this thesis and the work presented here. The presentation of these thoughts in such a manner should help to clarify the basic structure of the trust that will be formalised in the remainder of this work.

The chapter is separated into three distinct parts. The first part presents the thoughts of some of the major proponents of trust, from various disciplines. From these thoughts, it is possible to find a synthesis which describes to some extent the major workings of the concept of trust. Ordinarily, a work of this nature involves specialisation in one area which leads to a specific discussion about that particular area. For example, if the subject of study was authoring systems, it would concentrate mainly on these, writing from the point of view of, say, a Human Computer Interaction (HCI) researcher. A histogram of what might be covered would be that as shown in figure 3.1.

Trust, however, is all-pervading, as has already been discussed. What this thesis does is provide a wide-ranging overview of the subject of study, trust, with certain aspects of that subject covered more deeply than others. It is possible to split the subject of study across the various research boundaries, showing that the work covers a large amount of ground, with some subjects covered more extensively than others. Taking the fields of social psychology, sociology, philosophy, Game Theory and DAI for example, the area covered would be as shown in figure 3.2.

In the first example (figure 3.1), one subject has been addressed very deeply, perhaps touching on a few others. In the second (figure 3.2), a large number of subjects have been addressed, some more deeply than others. What is important for a thesis is not the depth, but the area covered by these histograms. Trust is

[4] For reviews of trust, see especially Golembiewski and McConkie's 1975 work, also the collection of essays in Gambetta, 1990. Mentions of trust abound in much of the literature pertaining to cooperation also, some of which are mentioned in this chapter.

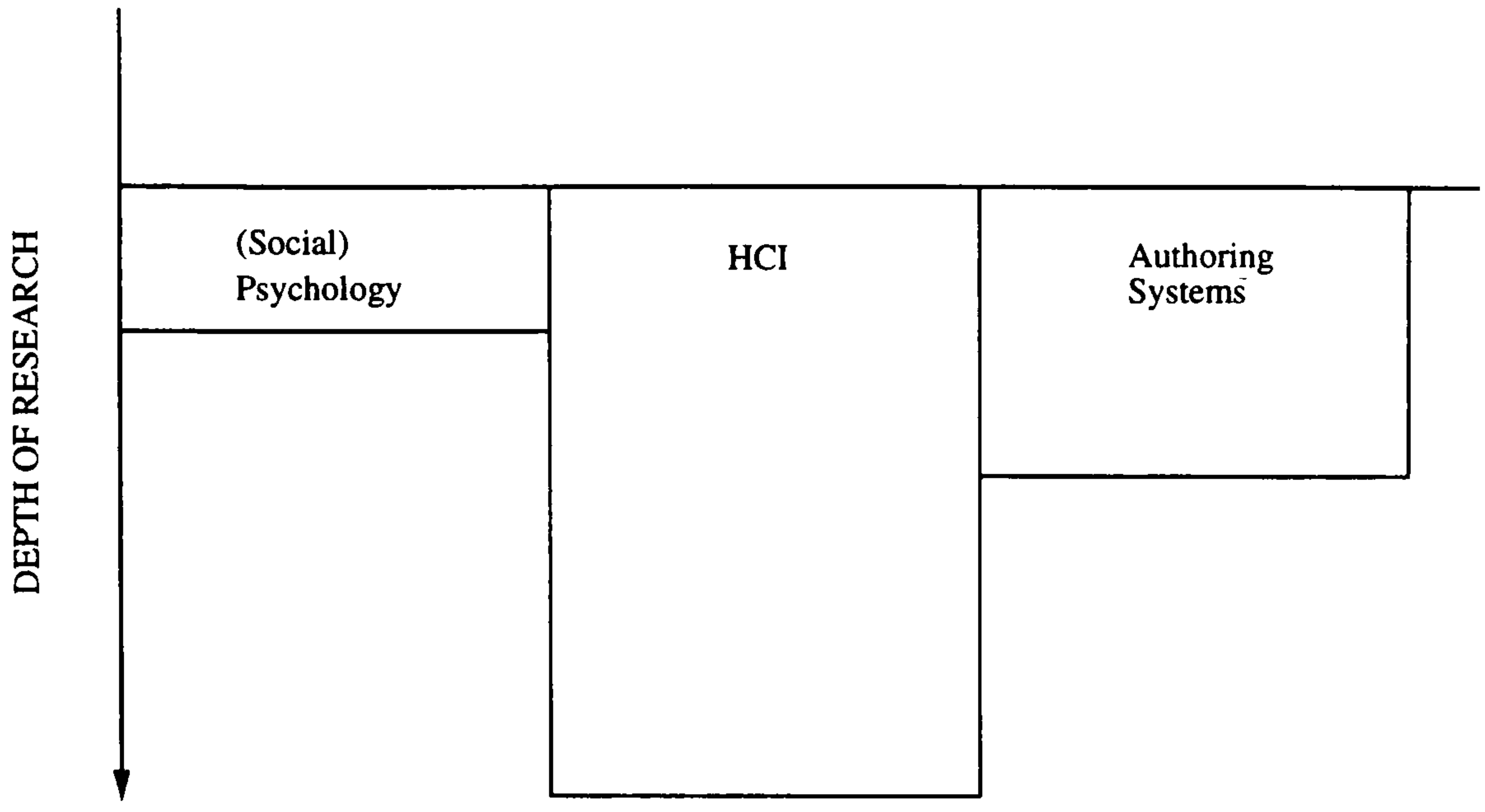

Figure 3.1: Histogram depicting possible research depth in one specific area.

unquestionably a multi-disciplinary subject. Thus, for trust, delving only into social psychology, for example, would be informative as regards what social psychologists think about trust, but would suffer from not showing a picture of trust *as it really is*. The concept of trust needs to be studied in this cross-disciplinary fashion. As a result, this chapter is at some times deeper than others, where the particular aspects of trust deserve it (from the point of view of the work as a whole). At the end of the first part of the chapter, there will have been presented a thorough overview of many of the contemporary views of trust.

The second part of the chapter amalgamates these views, determining similarities and differences between them, critically analysing their contribution to the general understanding of trust. In addition, other aspects of trust are presented, and the literature as a whole is discussed, particularly with respect to how it touches on the remainder of this thesis. For example, discussions of trust as an aid to the reduction of complexity, and trust as a commodity are presented. The aim of this section of the chapter will be fourfold:

- To confirm the existence of many aspects of trust.
- To verify that there are many views of trust that are possible or imagined, but that most of these views see trust as a major aspect of society and societal existence.
- To provide a deeper understanding of the place trust holds in society today, in

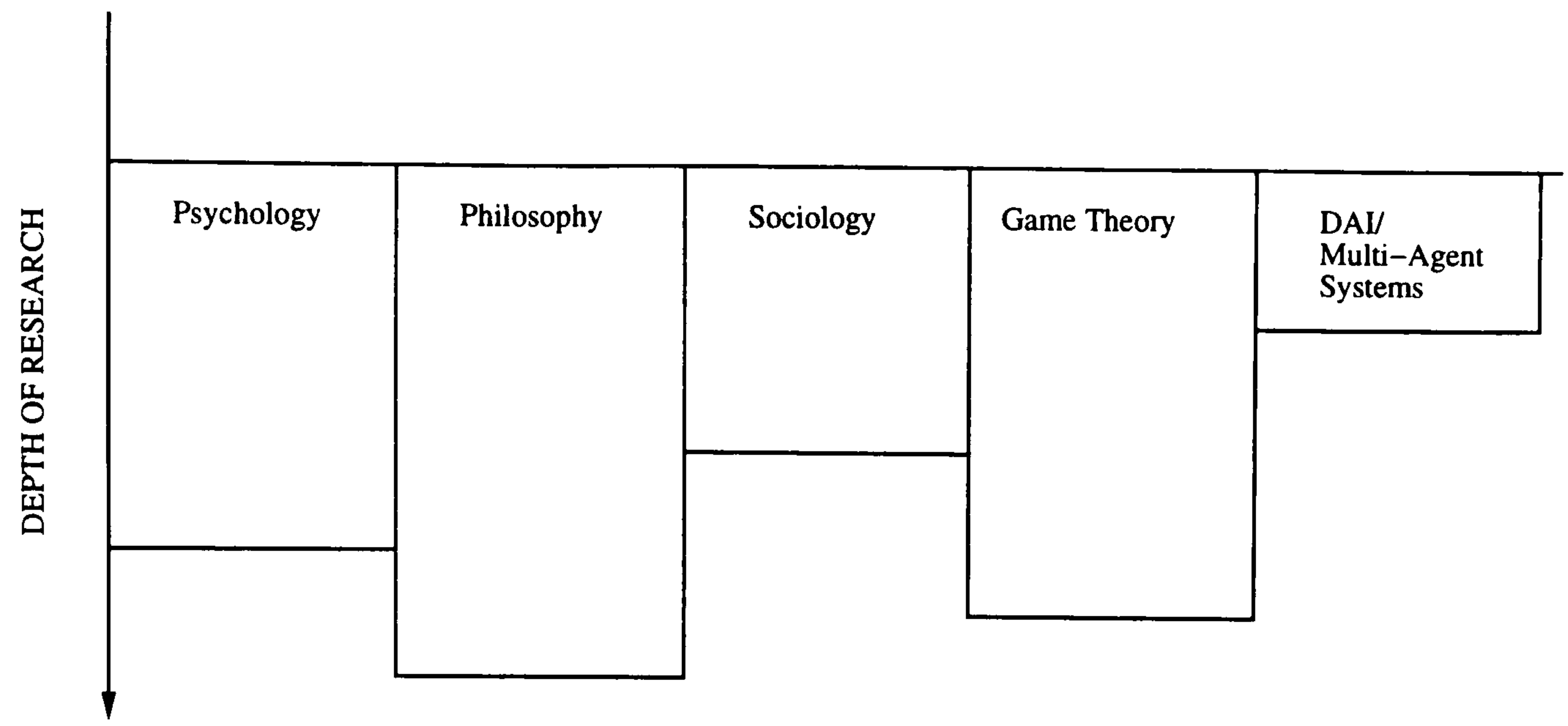

Figure 3.2: Histogram depicting possible research depth across several areas.

terms of its importance, and its sometimes invisible presence.

- To show that, despite common agreement as to the nature of trust and its importance, adequate definitions and conceptualisations are rare, vague, and not particularly useful, either in terms of understanding the concept or providing the benefits of that understanding to fields such as DAI.

This thesis is about trust, but it approaches the subject from a relatively novel angle. Firstly, it shows that, at present, although similarities between fields and definitions exist, there is an inherent vagueness about discussions of trust. Secondly, it attempts to clear away this vagueness by the use of particular tools which are available. In terms of the thesis, the tools of use are DAI as a modelling tool, and mathematics for the formalism. In addition to DAI being used as a tool for modelling, the thesis proposes that trust is of use to DAI as a field which is inherently social.

The final part of this chapter presents a discussion of the tools that can be used in order to attain a deeper understanding of trust. Firstly, it discusses the work of the mathematician George David Birkhoff in aesthetics and ethics, showing that what is proposed in this thesis is not a new idea, as far as attempts to formalise 'emotional' concepts go, and that, although Birkhoff's models were simple, they had an intuitive appeal to them. Taking the formalism developed in this thesis, the same applies. As Birkhoff states:

"I make no apologies for the simple character of the ideas involved, since it

> is inevitable that the initial results obtained be rudimentary. As possibly suggestive in this connection, it may be recalled that the first classification of matter as solid, liquid, or gaseous provided a crude trifurcation of nature, which ultimately led to the mathematical theories of elasticity and hydro-dynamics."
>
> Birkhoff, 1956, page 2200.

In other words, what starts out as a simple intuitive formalism will undoubtedly enhance our understanding of the concept, and may eventually, when built upon (if necessary) become the bedrock for the complete understanding and use of that concept. This thesis claims that the formalism for trust will provide a basis for the better understanding of trust.

The chapter goes on to critically analyse the research areas that have so far attempted to discuss trust, and to present discussions of the possible tools that we have for its formalisation and experimentation. For example, sociology has presented many discussions of trust, as has social psychology, but these fail from lack of concreteness. With the relatively new sciences of Artificial Life (AL) and DAI, we are given the possibility to attempt such discussions from new, exciting points of view. As will be shown below, however, AL proves limited, at least with respect to trust, whilst DAI does not.

The following chapters aim to present a view of trust in a formal sense.[5] They, to a large extent, speak for themselves, and justify their own approach, when combined with the methodology discussed in chapter 2. The use of DAI, however, needs more explanation. Thus far, the field is relatively youthful and vigorous, with much of its potential still to be discovered. It has been acknowledged, however, that DAI is a useful tool for modelling societies and aspects of those societies (Bond & Gasser, 1988), and lately, such research has been carried out with promising results (see, for example Drogoul and Ferber, 1992). The thesis extends it further, however, since we use DAI as a tool for examining the formalism we present, and commenting on its performance in implementation (see chapter 7).

The conclusion is that DAI is an excellent tool for the task of experimenting with implementations of trusting agents, due to its underlying sociality and flexibility. On

[5]The use of the word 'formal' may upset some. What is presented in the following chapters is, however, undoubtedly more formal that anything yet presented with regard to trust.

its own, trust as a field of study remains vague. Allied to maths to provide a formalism, it becomes understandable, and a serious object of discussion. With DAI to implement the formalism, it can be tested further to provide serious answers to serious 'what if' questions. In this respect, DAI is an excellent research tool, not only for trust, but also for other social phenomena such as morality[6] and ethics. DAI, on the other hand, is a thriving field of study which appears to have no need for an understanding of trust. The thesis argues that this is not the case, that DAI will benefit greatly from the implementation of trusting agents, since they will be more robust, less prone to errors of judgement, and consequently more efficient.

3.2 Part One — Insights into Trust

The literature available on trust is substantial, and there are some researchers whose work is of particular note, for differing reasons. The main thrust of work on trust in the past has come from three main areas, those of sociology, (social) psychology, and philosophy. Within these three spheres, some research efforts stand out in particular. These have been carried out by Morton Deutsch, Niklas Luhmann, Bernard Barber and Diego Gambetta. These researchers have done much to illuminate the problems associated with trust, and have published significant work in the area. Indeed, Luhmann and Barber in the sociological field have done much to solidify the conceptions of trust used in this thesis, as has Deutsch for his part in psychology. Bearing this in mind, it is useful to present their findings and propositions as independent units, with the aim of following their ideas through, before proceeding to more general presentations. The next few sections, then, present the thinking of these researchers.

3.2.1 Morton Deutsch

Perhaps the most popular and widely accepted definition of trust is that of Deutsch (1962), which states that:

> (a) the individual is confronted with an ambiguous path, a path that can lead to an event perceived to be beneficial (Va^+) or to an event perceived to be harmful (Va^-);

[6]Embedding morality into simple agents is, in fact, what Danielson has done (Danielson, 1990; Danielson, 1992b; Danielson, 1992a).

> (b) he perceives that the occurrence of Va^+ or Va^- is contingent on the behaviour of another person; and
>
> (c) he perceives the strength of Va^- to be greater than the strength of Va^+.
>
> If he chooses to take an ambiguous path with such properties, I shall say he makes a trusting choice; if he chooses not to take the path, he makes a distrustful choice.
>
> (Deutsch, 1962, page 303)

The use of the word 'perceives' many times in this definition implies that trust is a subjective, or agent-centred notion, one in which the choices that are made are based on subjective views of the world. This is of importance in the discussions and formalism to follow, and also accounts for much of what has been said earlier in terms of there being many views of trust.

If trust is based on individual perception, it is likely that in any one situation, different agents will see the situation differently. Consider, for example, a crime which several witnesses observe. Each witness will see things differently, noticing different details, and possibly wildly incorrect features about the happening. These differences come out in questioning later (Deffenbacher, 1991). The case with trust is similar, in that different agent's perceptions are different, thus, their estimate of costs (Va^-) and benefits (Va^+) will be different.

A second observation that can be made also concerns costs and benefits. The fact that they are mentioned at all is of importance — Deutsch is suggesting that trusting decisions are based on some form of cost/benefit analysis (Shapiro *et al.*, 1992; Dasgupta & Pearce, 1972), and in fact, this assumption is carried through by many other definitions of trust, in one form or another (Shapiro *et al.*, 1992; Golembiewski & McConkie, 1975). When we consider cooperation, the idea of cost and benefits becomes more important (Williams, 1990). There is, however, a problem, in that determining the costs and benefits of each individual outcome of a situation is inherently time-consuming; "One could spend hours analysing the costs and benefits of each situation in order to derive the maximum benefit from it. However, time is valuable too, and clearly the sensible approach to this problem of processing limits is to develop a scheme in which extensive intellectual work is only done under certain circumstances" (Good, 1990, page 42). Trust allows such limits to be addressed, as it allows the truster to assume that certain things are 'as given,' thus plans do not need to be made to allow for particular circumstances. This is what Luhmann means when he states that without trust we would not get out of bed in the morning (Luhmann, 1979). The environment is so complex that we have to trust in certain things in order to reduce that complexity.[7] Thus, when determining the costs and benefits of a particular route from a situation, certain assumptions can be taken 'on trust,' to risk a somewhat tautological explanation.[8]

Types of Trust

Many definitions of trust stem from Deutsch's work on the subject. His 1973 book expands the definition further, and presents clarifications, eventually arriving at the definition of trust as confidence, which is confidence that one will find what is desired from another, rather than what is feared (page 148). The notion of confidence crops up in much of the psychological discussions of trust, so, Scanzoni (1979) defines trust as "[an] Actor's willingness to arrange and repose his or her activities on [an] Other because of confidence that [the] Other will provide expected gratifications" (page 78). According to Rempel and Holmes (1986), it is "the degree of confidence you feel when you think about a relationship."

Much of the literature available refers to the idea of trust being confidence in some aspect or other of a relationship. Deutsch, however, presents many other aspects of trust in his 1973 work. To do so, he uses the story of *The Lady or the Tiger.* Briefly, the story is that a Princess has a suitor, who is discovered by the King. Unfortunately, the King is displeased, and so the suitor is thrown into a pit with two exits. Behind one exit is a ferocious tiger, behind the other a beautiful lady (presumably in lieu of the Princess). The young man is about to make a choice, when he notices the Princess pointing subtly to one of the doors. He immediately chooses that door. The reader is left to imagine the results. Does the Princess wish to see her love survive, even if he is in the arms of another woman, or would she rather send him to his death, given that choice?

[7]The reduction of complexity is discussed further below.

[8]An explanation which illustrates the dichotomy discussed earlier — we determine levels of trust using such mechanisms, which may themselves refer to 'black and white' trust.

Deutsch suggests that another question is as valid. Why does the suitor immediately choose the door the Princess points to? What doubts may he have about her intentions? This in itself leads to a question which Deutsch argues is central to trust — under which circumstances is one prepared to make a decision where the potential negative consequences outweigh the potential positive consequences?[9] Deutsch proceeds to suggest several different circumstances where such a choice could be made (see also Golembiewski and McConkie (1975)):

1. **Trust as despair.** This occurs when the negative consequences of not trusting, or of staying in the present situation, are so great or so certain that the trusting choice is made out of despair. For the suitor, then, staying in the present situation means certain death by execution, so choosing a door is the lesser of two evils.

2. **Trust as social conformity.** In many situations, trust is *expected*, and violations lead to severe sanctions.[10] The suitor will choose the door because he may end up socially ostracised or labelled a coward. I might lend a friend money even when I know the chances of repayment are slim, simply because there is no socially acceptable reason for refusing. Lack of trust may well destroy the friendship. This may be one reason why credulity is more morally acceptable than distrustfulness (Hartmann, 1932).

3. **Trust as innocence.** The choice of a course of action may be made upon little understanding of the dangers inherent in the choice. This innocence may be rooted in lack of information, cognitive immaturity, or cognitive defect (pathological trust).

4. **Trust as impulsiveness.** Inappropriate weight may be given to the future consequences of a trusting choice. Compare Axelrod's 'shadow of the future' here (Axelrod, 1984) which refers to the same thing in a different vein — behaviour

[9]Indeed, it makes little difference which exit the Princess points to if the suitor does not know what the signal means — 'don't go that way,' or 'do go that way' can both be signalled the same way. The suitor makes his own decisions about this, and indeed makes his final decision regardless. The intentions of the Princess are unimportant — what *is* important is the information the suitor deduces from the signal and his subsequent decision.

[10]Note that the converse is less true — credulity, or blind trust, is often seen as a problem in individuals (Dasgupta, 1990), but is not generally subject to sanctions.

in the present in terms of cooperation or defection in a Prisoners' Dilemma is weighted by the shadow of the future — how the agent expects to be treated in the future as a result of his actions in the present.

5. **Trust as virtue.** "Cooperative action and friendly social relations are predicated upon mutual trust and trustworthiness" (Deutsch, 1973, page 147). Thus trust is naturally considered a virtue in social life. See also the quotes at the start of this chapter — once it is realised that trust is indeed of such importance for society, its status as a virtue can only increase.

6. **Trust as masochism.** The suitor may choose to open the door expecting the tiger rather than the lady since pain and unfulfilled trust may be preferable to pleasure. Since people tend to try to confirm prior expectations (Rempel & Holmes, 1986; Boon & Holmes, 1991), they will trust negatively, and generally find their expectations satisfied.

7. **Trust as faith.** The suitor may have faith in 'the gods' such that he has no doubt that the lady is behind the door, or he may have faith in preordained paths which mean that whatever awaits is fated and thus to be welcomed. Having faith to a large extent removes the negative consequences of a trusting decision (Deutsch, 1973).

8. **Risk-taking or gambling.** If the potential gains of winning (the lady) are subjectively far greater than the potential losses from losing (meeting the tiger), even should the risk be great, the gambling suitor would be prepared to take that risk — life may not be worth living without the lady in any case, and so the risk is worth taking. It should be noted, though, that these estimates of chances and magnitudes of losses or gains are subjective, and thus some gamblers may take ill-advised risks, acting as though they are gambling when, in fact, they are trusting ill-advisedly.

9. **Trust as confidence.** Here one trusts because one has confidence one will find what is desired rather than what is feared. When Deustch uses the word fear here, he is implying that in order to trust, one must first take a risk. This analysis is agreed with by many of the researchers in this field, for example Boon and Holmes (1988) and Coleman (1990).

Of these nine, Deutsch concentrates on the last, trust as confidence. Trust is, then "strongly linked to confidence in, and overall optimism about, desirable events taking place" (Golembiewski & McConkie, 1975, page 133). Thus, the truster always hopes for the desirable outcome, whilst taking the risk that the undesirable will come about (Boon & Holmes, 1991). Clearly, with Deustch's definition, this is sensible, since the negative consequences outweigh the positive, and so one would not enter into the relationship unless one was hopeful (but see point 6 above). Luhmann (1990) suggests that "Trust is only required if a bad outcome would make you regret your action." (page 98). This seems at first sight to mean the same thing, but with deeper thought, suggests small differences in the two approaches — something which made you regret trusting in the first place is not necessarily more costly than it could have been beneficial. In addition, Golembiewski and McConkie (1975) state that "the loss or pain attendant to unfulfillment of the trust is *sometimes* seen as greater than the reward or pleasure deriving from fulfilled trust" (page 133, my emphasis). Clearly, Deutsch's statement that costs are greater than benefits to be gained is too strict in this case, and also in Luhmann's. The formalism presented in chapter 4 takes the view that the potential costs are not *necessarily* greater than the potential benefits.

Deutsch's Hypotheses

Having presented a definition of trust, Deutsch suggests some basic assumptions, rooted in psychology, which trusting humans would appear to obey, followed by a set of hypotheses which he aims to prove or disprove with practical experiments. A selection of the assumptions are as follows:[11]

1. Individuals tend to behave promotively toward things that are perceived of positive utility for them, and contriently towards negatively perceived things (persons, events, and aspects of themselves).

2. The potency of this tendency is related to:

    (a) strength of utility perceived;

[11] We do not include all of Deutsch's assumptions since some are not applicable here. The interested reader is referred to Deutsch's own work, particularly *The Resolution of Conflict*, 1973, chapters 7 and 8.

(b) perceived increase (decrease) in likelihood of event if the person acts promotively (contriently) to it;

(c) perceived probability of event after such promotive (contrient) behaviour;

(d) intrinsic utilities of activities involved in promotive (contrient) behaviour;

(e) perceived immediacy of event's occurrence;

3. People (and here, we include agents) are in their own ways 'psychological theorists' and tend to make assumptions such as the above about others.

As far as intelligent agency is concerned, these basic assumptions hold true, since agents should be rational. The rational agent will, for example, act to increase expected utility (Simon, 1955; Preston, 1961; von Martial, 1992) and to survive in the world (Hobbes, 1946) thus obeying assumption 2. Assumption 3, concerning the assumptions agents make of others, is perhaps more interesting in that it involves the agent making a model of another and the other's plans, and about the world. In situated action, the agent behaves in a completely reactive manner, thus models are not possible (Agre & Chapman, 1987; Chapman & Agre, 1987) and unnecessary. For Hobbes, everyone was in a state of 'war against all,' thus the only model that was to be made concerned the potential harm to be had from the actions of others (Hobbes, 1946, chapter 13, see also Kavka, 1983). In more deliberative systems, a knowledge of others is much more important (von Martial, 1992), particularly where coordination is required. Trust is one such model, or an aspect of a more detailed social model.

Deutsch presents 19 hypotheses relating to trust, which he tests in more practical psychological experiments. The thrust of these hypotheses concerns how the individual is likely to behave in relation to others before and after a trusting decision. Some of these hypotheses are relevant here, and are presented below. For the remainder, the interested reader is referred to Deutsch, 1973, chapter 7. In the following, we use Deutsch's notation, thus, an individual makes a trusting choice if, in a situation where Va^+ (positive valency/utility) is less than Va^- (negative valency/utility), the individual chooses an ambiguous path (see above). Deutsch's basic hypothesis is related to expected utility theory (see Simon (1955)):

Hypothesis 1: "Given that Va^- is stronger than Va^+, a trusting choice will occur if: $Va^+ \times S.P.^+ > Va^- \times S.P.^- + K$." Where:

S.P.$^+$ is the subjective probability of attaining Va$^+$, and S.P.$^-$ to that of attaining Va$^-$. K is a constant which Deutsch calls the 'security level' for the individual, and which differs between individuals. Thus, an agent will trust when he perceives that the probability of being let down by the trustee is small enough that the likelihood is he will benefit. The things that effect the subjective probabilities are numerous, and include past experiences in similar situations, the experiences of known others, and their opinions, and confidence in the ability to influence the outcome of the situation.

This hypothesis treats the positive and the negative aspects of a situation as linked, thus when S.P.$^+$ is high, S.P.$^-$ will be low. There are situations when both good and bad things can occur with equal likelihood, however, and Deutsch suggests that, if that were the case, a trusting choice would not be made. In a situation where temporal considerations are necessary, however, the subjective probability is not all that is considered:

Hypothesis 2: "The more remote in time the possible occurrence of Va$^-$ as compared with that of Va$^+$, the more likely it is that a trusting choice will be made." (page 153). Thus, immediate gains are likely to be sought after, even when the costs of the future are known — people smoke now, since they enjoy it, in the knowledge that it may adversely affect their future life. The future is 'discounted' with respect to the present (Pearce & Turner, 1990).

Hypothesis 3 concerns the seeking of evidence for what decisions have been made. The suggestion is that once a choice has been made, the truster will tend to seek evidence in favour of this choice (Good, 1990), and that the more difficult the choice was to start with, the more evidence is sought. Note that this is not entirely rational, since the evidence which is sought is confirmational evidence, not objective evidence. Since it is sought, it will be found, generally speaking, possibly even where none exists; so, in a romantic relationship, one partner may read into the other's actions something which is not there — "he's just buying me flowers because he wants to go to the pub with his mates tonight," for example (see Rempel and Holmes, 1986). Such a 'vicious circle' effect is doubly difficult to escape following a breach of trust, even if the partner is truly repentant (Rempel & Holmes, 1986).

Hypotheses 4 to 7 are concerned directly with the intentions of agents, and are thus of interest here. For example, it is of use to know the intentions of another (or at least to be able to estimate them) when entering into a relationship with them which might be detrimental or beneficial to oneself. **Hypothesis 4** states that "The stronger a perceived motivation underlying a given intention, the more reliable it will be perceived as being" (Deutsch, 1973, page 155). **Hypothesis 5** takes this further, stating that "The stronger a person's commitment to his intention is perceived to be, the more reliable it will be perceived to be." (page 155). Commitment, here, refers to the "desire to avoid not doing what one intends." There are many sources of intention with regard to actions, some of which are an altruistic or malevolent intention, an exchange intention (where something is wanted from the other person), and a conscience-directed intention, where the desire is to obtain approval. Some intentions cannot be satisfied without producing the actual behaviour (e.g., the altruistic or malevolent) whilst others do not have to be carried through, and can be accomplished by other means. Thus, **Hypothesis 6** states, somewhat loosely, that "When an intention must persist over time, through changing circumstances, an individual is likely to perceive another's intention as more reliable if the source of the intention cannot be satisfied by means other than the production of the intended behaviour than if it can be satisfied otherwise" (page 157). In other words, the means to the end are important as far as intentions are concerned.

Hypothesis 7 is concerned with the *focus* of an intention. Thus, a person may have as the focus of an intention the effect that his behaviour produces in the trusting person, or his own conscientiousness or his own behaviour. Thus, "An intention that is perceived to be focused upon producing certain effects in another person will be perceived as being more likely to result in such effects than if the intention is focused elsewhere" (page 157). This hypothesis, in other words, states that when we wish to be trusted, our focus is on producing trusting behaviour from those we wish to trust us. If this is the case, we are likely to be trustworthy. If the desired end is to go to the pub with our mates, however, the focus is not on the truster, and we, as the trustee, are likely to be perceived as less reliable.

3.2.2 Niklas Luhmann

The Reduction of Complexity

Niklas Luhmann's seminal work *Trust and Power*, was first published in German in two parts in 1973 (*Vertauen*) and 1975 (*Macht*). It was first published in English in 1979. Luhmann's approach to trust is sociological, as opposed to Deutsch's more psychological, practical approach (particularly using the Prisoners' Dilemma). Luhmann's main thesis is that trust is a means for reducing the complexity of society. He sees it as a "basic fact of human life" (page 4). The complexity of the world is, for Luhmann, a distinct problem for those agents who attempt to align themselves with it, or to adapt to it — "The only problem that does arise is the relation of the world as a whole to individual identities within it, and this problem expresses itself as that of the increase in complexity in space and time, manifested as the unimaginable superabundance of its realities and its possibilities." Thus, for the individual within this system, it presents itself as of a complexity so unimaginable as to inhibit adaptation. And yet, adaptation happens (Holland, 1975). There are reasons and methods for this. The basic method, which is accomplished by many different means, is to reduce the complexity of the environment to such an extent that adaptation can take place (Luhmann, 1979).

In simpler mechanisms, the organism "locates itself in a selectively constituted 'environment' and will disintegrate in the case of disjunction between environment and 'world'." (Luhmann, 1979, page 6). For human beings, however, the case is not so simple, since they have the knowledge and capacity to select their environment, perceiving themselves as being able to make decisions which change their environment, affecting self-preservation and other aspects of their life. An important aspect of this is the interactions with others who are similarly acting to select their environment, possibly affecting our own (Good, 1990). Thus society lends its own problems to increase the complexity of the everyday world. And so, "further increases in complexity call for new mechanisms for the reduction of complexity" (Luhmann, 1979, page 7). There are means of doing this, from Hobbes's ultimate political authority to utility theory. The point is, Luhmann suggests, that "in conditions of increasing social complexity man can and must develop more effective ways of reducing complexity" (*ibid.*, page 7). Moreover, "Where there is trust there are increased possibilities for experience and action, there is an increase in the complexity of the social system and also in the number of possibilities which can be reconciled with its structure, because trust constitutes a more effective form [than, for example, utility theory] of complexity reduction." (*ibid.*, page 8). Thus, trust is a means of reducing complexity, which is stable, socially acceptable, and effective.

Risk

In common with the majority of others researching trust, Luhmann sees it as a means of handling risk: "It becomes ever more typical and understandable that decisions cannot avoid risk. Such awareness of risk — the risks of technological development or of investment, of marriage or of prolonged education — is now a very familiar aspect of everyday life." (Luhmann, 1990, page 96). Indeed, "Trust ... presupposes a situation of risk." (*ibid.*, page 97).

Does this acceptance of the inevitability of risk weaken the concept of trusting at all? Luhmann, and others, argue that it does not, rather the acceptance of risk and the means, via trust, to cope with and assimilate it into the decisions we make, enables us to exist in the complex society which is around us. Without an acceptance of risk, and a means of handling that risk, we would be unable to face life, partly because it is too complex, partly because the risk of getting out of bed in the morning would be too great (Luhmann, 1979). Less frivolously perhaps, a knowledge of risk and its implications allows us to make plans for the future which take the risks into account, and thus make extensions to those plans which should succeed in the face of problems or perhaps unforseen events. Unfortunately, it is often difficult to predict the magnitude of these events, and risk analysis remains a problematic area (Zeckhauser & Viscusi, 1990).

Social Psychology and Sociology

As Luhmann (1979) points out, "Social psychology ... constantly attempts to reduce the social sphere to individual personality variables, which is why it is in no position to account for these facts [why trusting choices are made] very clearly. One of the first lessons of a theory of social systems is that very different personality systems can be functionally equivalent in social systems, so that social systems may to a certain extent be free from the personality processes of individuals." (page 9, note 15). Both social systems and personality variables are of importance. Trust is a social phenomenon (one trusts oneself, but one also trusts others). Indeed, if trust is in fact a method for coping with the freedom of others (Luhmann, 1979; Dunn, 1984; Gambetta, 1990a), then it cannot be anything but social. Trust, then, does not operate merely on a personal level, but also on a social level. Without both levels, we do not see the whole picture.

In fairness, Luhmann is guilty of ignoring the benefits of social psychology somewhat. In fact, it appears that, at the level of the individual, trust is one thing, but viewed at the level of society, as an emergent property of a properly behaving society, it is quite another, appearing to obey certain rules. It does, in fact, obey these rules (Luhmann, 1979), but since the society is made up of individuals, if every individual chose to change those rules, they would change. Thus, again, you cannot have one (social aspect) without the other (individual behavioural tendencies). The pathological behaviour of a single individual will hardly (most of the time) affect society, as far as trust in concerned, but trust nevertheless functions at the individual level. And so, individual personality traits affect it. Taking trust at the social level and formalising it, as we have done in this work, allows us to embed that trust in individuals (here, rational agents) and observe it, both at the individual level and at the social, collective level.

In summary, trust within society is an emergent property of the interactions of trusting individuals in that society. Studying the phenomenon at either level whilst ignoring the other leads to the inevitable loss of understanding of trust as a personal and a social concept.

3.2.3 Bernard Barber

Bernard Barber, dissatisfied with the loose usage of the word 'trust' in many walks of life, set about attempting to solidify and define the concept in his 1983 work, *The Logic and Limits of Trust.* In fact, this work concentrated on two aspects, firstly, an attempt to clarify the concept of trust, and secondly, an analysis of American society, with the initial question of whether or not it was a trustless, or distrustful, society. We are concerned with the former aspect here.

Barber's view of trust, as Luhmann's, is inherently sociological, and both in fact have roots in the work of Talcott Parsons (see for example Parsons, 1970). They are, then, functional accounts of the workings of particular aspects of society. For Luhmann, one of the functions of trust is to reduce the complexity inherent in the environment, thus allowing existence in, and adaptation to, that environment, preserving a functioning which is "both different from and in some respects *more stable* than its environment." (Parsons, 1970). Barber builds on both Parson's and Luhmann's work (although confessing he had not read Luhmann's work until he was well into his own formulation). As a sociologist, Barber leans towards Luhmann and away from Deutsch in viewing trust "predominantly as a phenomenon of social structural and cultural variables and not ... as a function of individual personality variables" (Barber, 1983, page 5).

Barber's view of trust is that it is an aspect of all social relationships. Moreover, it implies some form of expectation about the future (see Barber, 1983 , page 7. See also chapter 2 of Luhmann, 1979). These expectations are for the most part based on relationships and social systems in the world, and Barber presents three that "involve some of the fundamental meanings of trust" (Barber, 1983, page 9). These are:

1. expectation of the persistence and fulfillment of the natural and moral social orders.

2. expectation of "technically competent role performance" from those we interact with in social relationships and systems;

3. expectation that partners in interaction will "carry out their fiduciary obligations and responsibilities, that is, their duties in certain situations to place others' interests before their own."

When we consider that Barber's analysis is slanted towards examining the society in America, i.e., the social systems of, for example, government and the learned professions, these expectations become more sensible than if we consider trust in romantic relationships or friendships (see Rempel, Holmes and Zanna, 1985, Rempel and Holmes, 1986, Boon and Holmes, 1991). Thus, we expect professionals to behave in a responsible fashion towards us, neither 'blinding us with science' nor attempting to 'pull the wool over our eyes.' However, in a general sense, trust is an expectation that the natural, physical and biological order will continue to hold true (Barber, 1983, page 9). In a more specific sense, Barber asserts, trust is also placed in the moral social order. It is this which Luhmann refers to when he states that it is a basic fact of human life, and it is this that Bok (1978) says society will collapse without. It is this that is of relevance here. Of interest here is Garfinkel's work in the area, which involved some disturbingly effective experiments regarding the moral social order (Garfinkel, 1963).

As well as the moral social order, Barber suggests that there are two more specific types of trust. The first is the expectation of a technically competent role performance.

thus we trust our doctors to perform operations well, or we trust those we elect to govern the country in a sensible and efficient manner. The second type is the expectation of a fiduciary obligation and responsibility being fulfilled. That is, for those members of society who have moral obligations and responsibilities (to put the interests of others before their own), we expect that this will be done. The two are different, since although we can monitor the competence of those in most professions (by their results, if nothing else), we are frequently in a position where we know much less than they about what they are doing, and thus must trust them not to use this power against us, but for us, with our well-being in mind.[12]

The question as to whether trust can be transferred or generalised across relationships and between systems is addressed by Barber, who states that "It should be an axiom of social analysis that actors who perform competently or show great fiduciary responsibility in one social relationship or organisation may not necessarily be trusted in others... Trust cannot necessarily be generalized." (Barber, 1983, pages 17–18). In this work, however, we have ventured to suggest one method, not for generalising over systems, but between relationships, such that, in the same situation, the same actor can be trusted, to some extent, dependent on his behaviour towards others. Clearly, a doctor builds a reputation based on his competence and fiduciary responsibilities. This reputation *is* generalisable to other patients, but not necessarily to other roles (e.g., as a family man). See chapter 6 for a discussion of this.

Trust, then, according to Barber, has three meanings, one very general (trust in the moral social order) and two more specific (technical and fiduciary). At any time, we may place all three in one person. The functions of trust, therefore, include the provision of a social ordering, "providing cognitive and moral expectational maps for actors and systems as they continuously interact." (*ibid*, page 19). This is of importance for agents in DAI — see below. Another function of trust, Barber asserts, is social control, specifically with regard to technical and fiduciary trust. For this, Barber states that social control has a positive, rather than the usually understood negative, meaning — "the mechanism for providing the necessary means and goals for the achievement of social system requirements" (*ibid*). For this to work, power is necessary, and in order to place power in the hands of, for example, the government,

[12]To that end, professional societies and the law can come in handy — see sections 4.7.5, 5.3 and 8.5, see also below in this chapter, particularly section 3.3.3.

trust is necessary (Dasgupta, 1990). "Thus the *granting* of trust makes powerful social control possible. On the other hand, the *acceptance* and fulfillment of this trust forestalls abuses by those to whom power is granted" (Barber, 1983, page 20). At least, in theory; as Dasgupta notes, in order to trust those in power, it is necessary to believe that they can be removed from power (Dasgupta, 1990), through the ballot box or otherwise. The withdrawal of trust is sometimes apt to be violent, another indication of its importance.

3.2.4 Diego Gambetta

Diego Gambetta's collection of works under the title *Trust* (Gambetta, 1990c) gathers together the thoughts of many diverse areas, from biology to economics. Of particular interest in this section is the final chapter of that collection, by Gambetta himself, *Can we Trust Trust?* In it, he presents a view of trust similar to that in the formalism presented in the following chapters. (The paper was found some way into the work documented here, but the similarities are often striking.) It is reviewed here for two main reasons:

- It presents a view of trust which, although from a different viewpoint than that presented here, echoes much of the views presented in this work as regards the workings of trust. It is thus of major interest both due to these similarities and because of its different viewpoint.

- It summarises a volume, edited by Gambetta (Gambetta, 1990) whose theme was trust. It is thus of great applicability to this work, and is a major contribution to the trust literature.

Gambetta's view of trust, seen from the viewpoint of an amalgamation of differing viewpoints, much as in the present work, is similar to the present work in many important ways. The most important similarity concerns the use of values. As discussed in chapter 2, the use of explicit values for trust can be seen as something of a problem, particularly since trust is subjective enough that the same *value* may mean different *levels* of trust for different agents. However, the use of values does allow us to talk succinctly and precisely about specific circumstances in trusting behaviour. In addition, it allows the straightforward implementation of a formalism. Gambetta uses values in the range 0 to 1. In other words, trust is a probability, which he defines as follows: "trust (or, symmetrically, distrust)[13] is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it[14]) *and* in a context in which it affects *his own* action." (Gambetta, 1990a, page 217).

As the final part of this definition suggests, trust is a means of coping with the freedom of others, and how this affects us (Luhmann, 1979; Dunn, 1990). In other words, trusting a person means that the truster takes a chance that the trustee will not behave in a way that is damaging to the truster, given that choice.

Gambetta's definition excludes certain aspects which are of importance to trust, however, in that clearly it refers only to trusting relationships between agents, not, for example, between agent and environment. It also excludes those agents whose actions have no affect on the decision of the truster, despite trust being present. The first circumstance is more problematic since we do have some form of trust in the natural order of things (Barber, 1983; Luhmann, 1979). However, from the point of view of the present work at least, this trust is simply implicitly assumed — that walls don't move, or more practically that if an inanimate object is moved, it stays where it is put, all other things being equal. The second exclusion, of those whose actions do not influence the decision, is less of a problem, since, as trust is based in the possible future actions of those the truster perceives will be able to affect him, having trust in someone who is not perceived to be able to affect the truster does not affect the decision that is made. In other words, an agent may or may not trust his bank manager to look after his money, but this has little effect on the decision to trust his office mate to post a letter for him. Different situations, such as the decision to invest money, for example, would naturally concentrate on the agent's trust in his bank manager.

Of course, there is always a risk involved in trusting (Luhmann, 1979; Luhmann, 1990), and thus people tend to try to remove the need to trust by establishing pre-commitments, or *constraints*, on the truster and the trustee. In a discussion closely related to those presented later in this work on legal aspects of trust, Gambetta states that pre-commitments establish that we as trusters can also be trusted, thus we make

[13]As Luhmann notes, trust and distrust are not merely opposites, but are functional equivalents of each other (Luhmann, 1979; Luhmann, 1990).

[14]This corresponds directly to the notion of fiduciary trust for Barber, see above.

promises, sign contracts, and so forth. Ulysses, when he tied himself to the mast of his ship, showed a lack of self-trust, but combatted this by committing himself. Because he could not trust his sailors, he made sure they could not hear the Sirens' songs or his orders to take him closer to the Sirens (Elster, 1979). Such a form of pre-commitment is usually applied towards others, so that they may rely on, or trust, us. On a weaker note, contracts and promises are made, not physically binding, but costly to renege upon.

An interesting point in Gambetta's paper concerns competition. It is accepted (for example, in economics, concerning monopolies) that there are times at which we would say that cooperation was not a good thing, and would wish to discourage it. One example is monopoly control in the business sector, another is the control of criminals — a cooperating society of robbers is quite undesirable, where cooperating police forces are unquestionably useful in this context. It seems, then, that we should try to find "the optimal mixture of cooperation and competition rather than deciding at which extreme to converge." (Gambetta, 1990a, page 215). With regard to competition, even then, cooperation is of great importance, since, "Even to compete, in a mutually non-destructive way, one needs at some level to *trust* one's competitors to comply with certain rules." (*ibid.*, page 215). As Gambetta notes, "there is a difference between outdoing one's rivals and doing them in" (*ibid.*), and inter-species rivalry is considerably more inclined towards the former.

Despite the important insight of using values for trust, Gambetta does not develop the idea in any concrete fashion. He does mention contracts and how the law and legal aspects will lessen the need to rely on trust (pages 221–222; see also chapter 8, section 8.5 in this work). In addition, the concept of a threshold is mentioned at several points, and closely resembles the cooperation threshold discussed in chapter 4, often with significant insight: "We may have to trust *blindly*, not because we do not or do not want to know how untrustworthy others are, but simply because the alternatives are worse." (page 223). This is a direct similarity to Deutsch's view of trust as despair (see earlier in this chapter). Deutsch ignores this aspect of trust, but Gambetta includes it not as a separate aspect, rather as a part of the whole which is necessary in order to get the 'whole' picture of trust. In the present work, trust is seen as somewhere between the two — clearly, trust as confidence (to use Deutsch's definition) is a major aspect of proper functional trust, but the other, less important aspects allow an agent to behave in a reasonable manner even when rationality fails. It is this which trust as despair actually allows, and this is why Gambetta refers to it in such a way.

The final part of Gambetta's paper concerns whether or not trust is a rational and sensible option, i.e., whether we can trust trust, and correspondingly distrust distrust — that "it can be rewarding to behave *as if* we trusted even in unpromising situations" (page 228). So that, given trusting behaviour, others will learn to cooperate, much as in Axelrod's Prisoners' Dilemma tournaments, where Tit for Tat actually encouraged cooperation. And even Tit for Tat needed at first to behave as if it trusted, so that it cooperated on the first move. There are some conclusions to present here:

- "I cannot will myself to believe that X is my friend, I can only believe that he is." (page 231). In other words, trust cannot be brought about at will. Indeed, the statement 'trust me' does not work unless trust is present in the first place (Boon & Holmes, 1991).

- "if X detects instrumentality behind my manifestations of friendship, he is more likely to reject me and, if anything, trust me even less" (page 231). This is of particular importance in the formalism to be presented, and is similar to Deutsch's idea of the *focus of intention* (see earlier, also Deutsch, 1973). If the workings of the formalism are known, deceit can become problematic. Bearing this in mind will help armour the trusting agent against such actions.

- Trust is of use in situations of ignorance. However, the seeking of evidence in situations often affects the evidence itself (Boon & Holmes, 1991). Thus, "While it is never that difficult to find evidence of untrustworthy behaviour, it is virtually impossible to prove its mirror image." (Gambetta, 1990, page 233). In other words, once distrust has set in, it is particularly difficult to know if such distrust is justified since such experiments will not be carried out. Trust is capable of spiralling dramatically downwards (Golembiewski & McConkie, 1975). Conversely, however, it is capable of spiralling upwards, and being self-reinforcing (*ibid.*).

Perhaps most importantly from the point of view of this work, and multi-agent systems, Gambetta states that "sustained distrust can lead only to further distrust. Trust, even if *always* misplaced, can never do worse than that, and the expectation that it might do at least marginally better is therefore plausible." In other words, a knowledge of the workings and usefulness of trust can help artificial agents get along better in the unpredictable world they exist within.

3.3 Part Two — Generalities of Trust

3.3.1 Risks, Costs, and Benefits

A decision to trust is a decision laced with risk (Rempel *et al.*, 1985; Boon & Holmes, 1991; Luhmann, 1990). All of the literature on trust appears to agree with this statement, to a greater or lesser extent. According to Boon and Holmes (1991), the taking of risk is necessary in ongoing relationships, both in order to confirm the trust that already exists, and to strengthen it, should the risk be justified. On the other hand, if the risk is not justified and the other party 'defects,' trust is revised downwards dramatically. As Govier (1992) states, "Trust is often founded on evidence, but even when our expectations are well grounded there is an element of risk in trust, a chance that those whom we trust will not act as expected" (page 17). In Deutsch's example of the *Lady or the Tiger*, the suitor takes the risk that he will be killed, whilst trusting the Princess to choose what she will for him.

Despite its usefulness in both human and artificial senses, trust remains a risky proposition, and one which can be incorrectly placed: "However indispensable trust may be as a device for coping with the freedom of others, it is a device with a permanent and built-in possibility of failure" (Dunn, 1990, page 81).

The estimation of that risk remains a problematic area (Zeckhauser & Viscusi, 1990), and frequently is virtually impossible. Boyle and Bonacich introduced a simple method for risk estimation in 1970, using the Prisoners' Dilemma matrix for an example. Given the matrix:[15]

		B	
		c	d
A	c	x x	z w
	d	w z	y y

[15]See appendix A for clarification.

Where $w > x > y > z$, the risk inherent in making a cooperative decision was estimated by $r = (y - z)$, contrasted with the possible gains from cooperation of $g = (x - y)$. The higher the risk, the less likely is cooperation. The temptation to defect, $t = (w - x)$ also plays a part here, so that the higher it is, the less likely is cooperation, and since this is common knowledge, the less trust there is. Boyle and Bonacich finally suggested an intuitive 'caution index,' taken as $\frac{\sqrt{r \cdot t}}{g}$. This is because we can standardise r and t using g, giving caution as the geometric mean of the two, since the motivation to cooperate increases as g increases. They define trust as the "difference between Player's expectations of the probability that the Opponent will cooperate and the degree of cooperation implied by the caution index" (page 130). This is simplistic, but according to Boyle and Bonacich, is upheld by various experimental simulations. So, the higher the caution index (the lower g), the less likely the agent is to cooperate, or trust.

As useful as the Prisoners' Dilemma is, it is limited (see the discussion on Game Theory below, also chapter 7). Estimating a caution index in situations where everything of relevance is known or can be easily found is a fairly simple affair. It is much more difficult to estimate risks under conditions of uncertainty, or even where judgement is clouded, for example when the decision is a matter of life or death, or if the risk is far into the future and is thus discounted (Pearce & Turner, 1990). From Boyle and Bonacich, and indeed from much other literature, the risk can be said to be intimately related to the amount of perceived costs and benefits in a situation — the higher the potential costs, the higher the risk, with the amount of benefit having a more or less equal effect on risk, so that, in a simple estimate, risk = costs/benefits (Marsh, 1992).

Whatever the relationship between risks, costs, and benefits, the relationship between risk and trust is clear and acknowledged. The taking of the risk either reinforces the trust that is there already if cooperation ensues, or if trust was not there before, builds trust if cooperation ensues. If cooperation does not ensue, then the risk of trusting is shown, and the trust decreases accordingly. If it was high initially, and the risk of rejection was great, then rejection will cause a large loss of trust (Boon & Holmes, 1991), which is particularly hard to rebuild since the truster will tend to look for evidence which proves that trust is not warranted (Good, 1990).

3.3.2 Confidence, Familiarity, and Faith

To some extent, this section presents a discussion of what trust is *not.* With respect to confidence, clearly trust and confidence are particularly closely related in that they are both expectations of things which can be disappointed. As Luhmann points out, however, the difference between the two is that trust presupposes an element of risk, whereas confidence does not (Luhmann, 1990). Nevertheless, trust and confidence can be routine behaviours. The distinction is more complex than this. In a situation of confidence, alternatives are not considered — leaving the house without a gun every morning is a sign of confidence that the weapon will not be needed. If alternatives were considered, however, and you still left unarmed, then the situation would become one of trust (Luhmann, 1990). The choice under confidence may be disappointed, but if trusting choices are made and disappointed, then the truster will regret the decision to trust. This may be the case, but the truster should not be blamed for trusting, rather the trustee should be blamed for reneging on that trust (Hertzberg, 1988). As an example, leaving the house without a weapon and getting shot or mugged is not our fault, even if we considered the situation. Relying on not getting mugged is our fault, however, if we go into a rough area: "trust can only concern that which one person can rightfully demand of another" (Herzberg, 1988, page 319), and since we *know* the area is rough, we cannot demand anything of anyone.

3.3.3 Trust as a Commodity

Trust has been looked at from many angles, but a novel and important one is that of Economics. As Dasgupta (1990) states; "Trust is central to all transactions and yet economists rarely discuss the notion." (page 49). From the point of view of entering into transactions with strangers, this statement is extremely pertinent. Questions such as 'is the good merchantable,' 'is the dealer reputable, not fly-by-night,' 'where can I take it if it goes wrong,' and so on, have to be asked. A degree of trust is necessary to enter into such transactions (except when there is no choice — as in trust as despair (see earlier)). Dasgupta notes that there are ways around having to trust too much, notably the presence of enforcing agencies to ensure our safety as consumers. In other words, the incentives to be untrustworthy for the dealer must be large enough that trustworthy behaviour is more likely than untrustworthy behaviour (and this is all we can hope for). There are several points which Dasgupta raises:

1. Absence of punishment means that the incentives to be trustworthy are absent. Since this is common knowledge, no-one will enter into transactions;

2. The threat of punishment must be credible, otherwise the threat does not exist. In other words, the enforcement agency (government, central or otherwise; or professional institution, for example) must be trustworthy, and seen to be trustworthy;

3. Following from 1 and 2, trust between persons and agencies is interconnected. So, if you lose trust in the enforcement agency, you will not enter into an agreement. Likewise, you *will* lose trust in the enforcement agency (e.g., the government) if it cannot be punished (voted out of power) itself;

4. You do not trust someone to do something simply because he says he will do it. This is an important point, implying that 'blind trust' is ill-advised (and, as we argue in the next chapter, not really trust at all). You trust because of perceived dispositions, abilities, etc.

5. Since you have to look at the dispositions of the trustee, it follows that when you trust, or consider trust, you look at the world from the perspective of the trustee when he has to comply with the agreement.

6. Trust has no measurable units, but its value, its worthwhileness, can be measured. It is, thus, a commodity, like information and knowledge.

7. Dasgupta takes trust to be "correct expectations about the *actions* of other people that have a bearing on one's own choice of action when that action must be chosen before one can *monitor* the actions of those others." (page 51). It is, in this sense, a way of dealing with the freedom of others (see especially Luhmann (1979,1990)).

Much of what is stated here has been reinforced elsewhere. One point which is important is that of trust as a commodity. Dasgupta is not alone in holding this view. Baier (1986) shares it, as is shown by the quote at the start of this chapter. Indeed, trust is seen by many to be a common good, and a good which paradoxically decays with misuse, and grows with use. The more trust we exercise in others, the more trust is generated. If it is common knowledge that everyone is trusting and trustworthy, trust itself is strengthened. The notion of trust being a common, or social good, is an especially pertinent one, since it highlights many of the problems we may have in discussing the phenomenon. Because it is, by and large, invisible and accepted, any attempt to 'pin it down' is doomed to failure — there is always something missed. Thus, previous attempts to define trust, from within the narrow confines of specific fields of research, have failed to notice other aspects of the phenomenon. Undoubtedly this attempt will also fall somewhat short of the mark, but as is mentioned elsewhere, this is to be expected, and in any case the formalism can be refined to take this into account.

As Bok (1978) states, trust is a social good which has to be protected. Somewhat paradoxically, the protection for trust appears to be its use. Thus, the more we trust, the stronger it becomes, and with it, the society which is built around it (Bok, 1978; Baier, 1986; Yamamoto, 1990).

The notion of trust as a commodity does have its problems, however; not being able to measure trust in units is in itself an obstacle to our understanding of it. However, we are able to measure its *value*, in terms of what benefits it brings to us. In the following chapter, we present a view of trust as measurable — a person trusts another by a certain amount — there are no units, as Dasgupta states, and the interpretation of the values is left to the reader. For the purpose of this thesis, however, the values represent subjective probabilities of trustworthy behaviour on the part of the trustee.

3.3.4 Variable and 'Absolute' Trust

Much of what has been discussed so far mentions, but inevitably glosses over, the problem of values of trust. In the previous section, for example, we mention that trust has a value which has no units, but can still be measured in terms of such vague notions as 'worthwhileness' and 'intrinsic value.' In other parts of this chapter and thesis, we mention the idea of trust being an absolute medium — one either trusts, or one does not.

Rather than being problematic, this suggests that trust has certain threshold values, above which it is possible to say that something or someone 'is trusted,' below which it is possible to say that it 'is not trusted.' As Gambetta notes, this threshold is different not only for different people, but also for the same people in different circumstances (Gambetta, 1990a. See also Marsh, 1992). Then again, not trusting someone is not the same as distrusting them (Marsh, 1992; Marsh, 1994b), and it would be sensible to assume that there is also a negative threshold, above which one could say that something or someone 'was not trusted,' below which one could say that they 'were distrusted.' Figure 3.3 illustrates this.

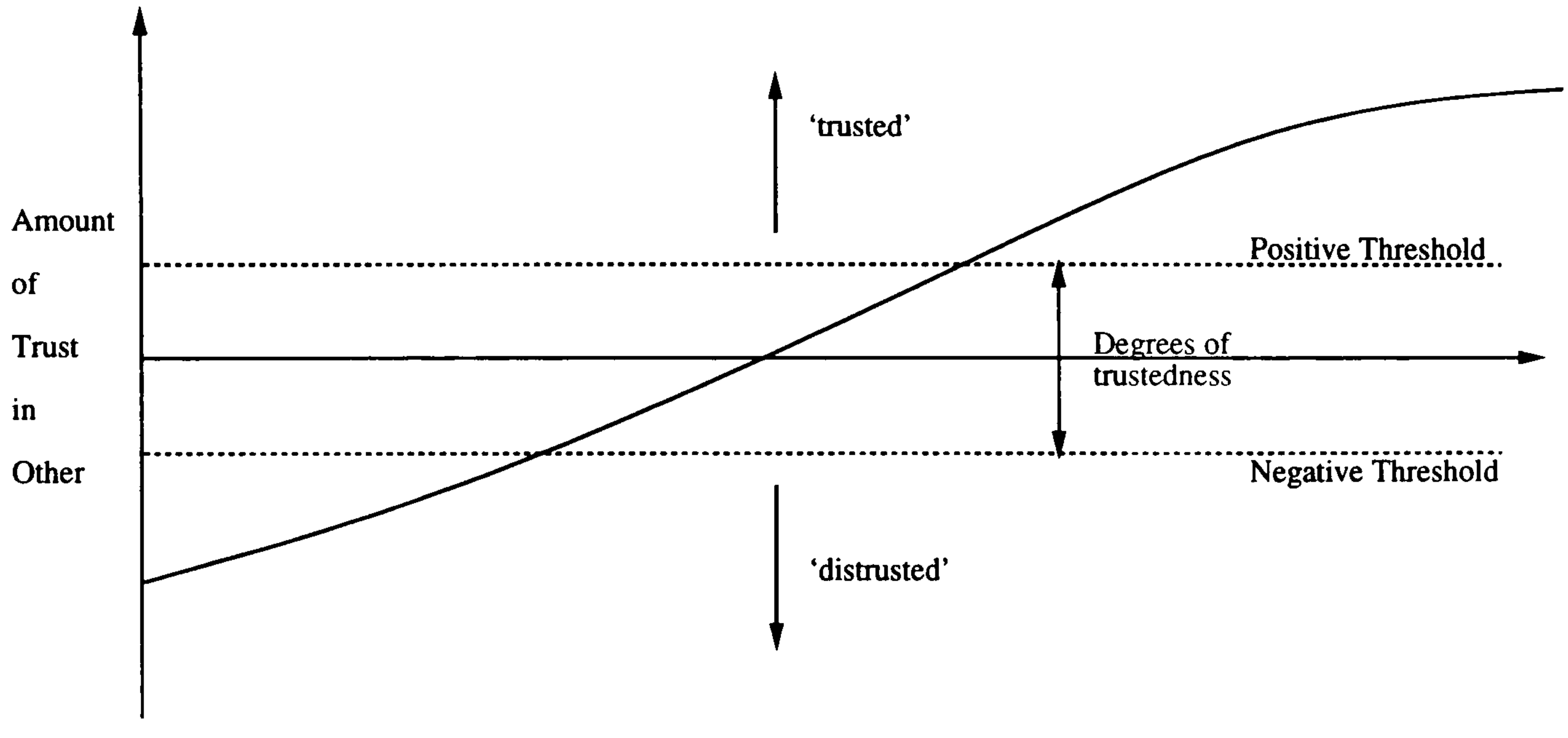

Figure 3.3: Positive and negative thresholds for trust.

3.3.5 Expectancies — Generalised Probabilities

Generalised expectations about others also tend to be suggested as part of trust by many, particularly Rotter, who states that trust is "a generalized expectancy held by an individual that the word, promise, oral or written statement of another individual or group can be relied on" (Rotter, 1980, page 1, see also Rotter, 1967 and 1971). Rempel *et al.* also touch on this when they define trust as "a generalized expectation related to the subjective probability an individual assigns to the occurrence of some set of future events." (Rempel *et al.*, 1985, page 96).

Referring to trust as a generalised expectancy implies that trust is an adaptation, a generalisation, or, as Luhmann states, a means for the reduction of complexity. Adaptation in the world is of critical importance for independent agents: for example, adaptation (via evolution) provides organisms with a means of becoming increasingly fit with respect to the environment, and it is this which enables organisms (agents) to learn from experience in beneficial ways (Holland, 1975). Whether or not trust is an evolutionary adaptation may be debatable (Gambetta, 1990a), nevertheless, it is a good generalisation across experience. In other words, trust enables agents to make general assumptions about situations in which they are ignorant of something, be it other agents, outcomes of previous situations, or in situations in which they lack some information. These decisions are most likely not as 'good' as well-informed decisions, in the sense that they may often result in an outcome within which the payoffs received by agents (the truster in particular) are not as much as they would have been under conditions of perfect information. This imperfection notwithstanding, the decisions that are made can be better than no decision being made at all, a state of deadlock, or worse. Trust, in the sense that it allows agents to make generalisations, is a powerful tool to the agent in a state of imperfect information.

3.3.6 Trust and Cooperation

> Cooperation requires trust in the sense that dependent parties need some degree of assurance that non-dependent parties will not defect.
>
> Williams, 1990, page 8.

The previous sections of this part of the chapter have introduced aspects of trust which are generally agreed upon by those working in the area. In addition, they have referred at several points to 'intelligent agents.' One of the claims of the present work is that trust can help artificial intelligent, or pseudo-intelligent, agents to reason about others and their environment, thus becoming more robust in situations of uncertainty, for example. Another is that having a knowledge of trust will help such agents in their considerations of cooperative situations, in conditions of imperfect information, or where the situation depends on speedy, accurate decisions, as quite often will be the case (consider the hypothetical example of agents in a flight control system for a jet airliner, for instance, where speed of decision is matched by the need for accuracy and robustness).

This section considers the effect that trust can have on decisions regarding cooperation with other independent agents, in cooperative and non-cooperative environments, and shows how one of the claims of the present work is justified with respect to intelligent agents; namely, as presented above, that trust can help independent agents in their decision making process with regard to cooperation. In order to do this, it argues that a climate of trust can encourage cooperation, and that trust allows cooperation when it would be unlikely were the trust not present or considered.

Cooperation is undoubtedly a good thing for society as a whole (Argyle, 1991). Indeed, it is questionable that society would exist without some cooperation, between governors and governed, police and public, car drivers, and so forth. As Argyle argues, cooperation is central to the whole existence of humans and humanity. Cooperation is not limited to humans either. In the animal world, cooperation exists between chimpanzees (Trivers, 1985), bats (Harcourt, 1991) and whales (Trivers, 1985). It appears that cooperation is, in evolutionary terms, a successful strategy, one which ensures the survival of the 'genes for cooperation.' Cooperation appears to be necessary for the very survival of the animals involved (Argyle, 1991). In humans, this is less obvious, but society depends on its presence.

Argyle (1991) gives three main ways in which people can be cooperative:

1. **Cooperation towards mutual rewards.**

2. **Communal relationships.** For example, marriage and love. Situations where altruistic concern is of importance. Also friendship, which, as well as providing its own benefits, sustains social support also.

3. **Coordination.** Coordination is necessary for all social situations, even if just to avoid bumping into others, for example when driving, or to hold a coherent conversation, or to work towards a group project. Clearly, cooperation is needed here.

All of these are important, and all are of some relevance to the present work. In fact, splitting the concept of cooperation in such a fashion does little to clarify it, particularly with regard to multi-agent systems within which artificial agents play a part. In addition, the problems of emotional attachment and potentially altruistic feelings on the part of an agent are beyond the scope of this thesis, although they are interesting avenues in themselves. Nevertheless, they illustrate the scope and importance of cooperation in society.

There have been in the past experiments with cooperation among small groups (face to face). Once again, Deutsch is a leader in the field. His view was essentially based on the work of George Herbert Mead. Deutsch states that "cooperation breeds new motives, attitudes, values, and capabilities." (Deutsch, 1962, page 275). As early as 1949, Deutsch developed a theory of cooperation and competition (Deutsch, 1949a; Deutsch, 1949b), and associated experiments to attempt to justify his hypotheses.

Indeed, this work bears a resemblance to the work Deutsch did later with trust and cooperation. In common with other researchers, many of Deutsch's hypotheses concerned groups and their feelings towards others (McNeel *et al.*, 1974). Thus, members of cooperative groups perceive themselves to be interdependent with that group, particularly if it is in competition with another (Argyle, 1991). Whilst acknowledging the importance and undoubted benefits of competition (Galliers, 1989), it is safe to assert that cooperation is often a beneficial strategy, particularly for those cooperating.[16]

Taking such benefits into account, the question remains as to where trust plays a part in the initiation and maintenance of cooperation. At the most intimate level, for example of marriage, trust plays a large part in the relationship (Boon & Holmes, 1991). In such a relationship, partners continually place themselves in the others' hands, with the knowledge that the other is free not to cooperate. Boon and Holmes (1991) give an example of a wife, tired after a long day's work, in the dilemma of whether or not to ask her husband to cook the evening meal. Her actually asking places her in his hands, with his different responses either justifying her decision to trust or not. As Deutsch states, "the initiation of cooperation requires trust whenever the individual, by his choice to cooperate, places his fate partly in the hands of others." (Deutsch, 1962, page 302). For the wife, the decision to cooperate is made when she asks her husband to cook. It follows that in order to initiate cooperation, risks must be taken (Rempel *et al.*, 1985; Rempel & Holmes, 1986; Boon & Holmes, 1991; Shapiro *et al.*, 1992). But the benefits are often worth the risk — the wife may well find that her husband is only too happy to cook the tea, with the associated benefits of not only a rest in the evening, but also a bolstering of her trust in him, and thus an increase in the likelihood she will ask another time. Trust often spirals upwards in this fashion (Boon & Holmes, 1991).[17]

[16] As Argyle (1991) and Gambetta (1990) argue, some cooperation is definitely not a good thing, for example between aggressors in a situation of war (Hinde & Groebel, 1991b) or between the members of an oligopoly, which indeed must be controlled by government. There are, then, situations where we would wish to discourage cooperation. As another example, consider how beneficial cooperation between criminals may be (Gambetta, 1990a).

[17] However, it also can spiral downwards (Golembiewski & McConkie, 1975). Consider the attendant losses for the wife if her husband refuses — not only does she have to cook (in a simple argument), but also she loses face and trust in her husband. Following this, she is not likely to ask again, since she *expects* a refusal, and her trust decreases again, even though she did not ask. In such situations, a risk *must* be taken to get the relationship back on the right footing, but such a risk becomes less

In less intimate relationships also, trust plays a part in the initiation and maintenance of cooperation (Golembiewski & McConkie, 1975; Deutsch, 1962; Deutsch, 1973). Also in business relationships, trust allows increased performance since it provides the benefits of "reduced need for monitoring behaviour and greater speed in making decisions" (Shapiro *et al.*, 1992, page 365).

In terms of the present work, there is much of interest in what has already been discussed. The formalism presented in the following chapters has ambitious ideals. One is that trust, or an understanding of it, can be instilled in artificial agents to provide robustness in the face of uncertainty and an acknowledged means of coping with the complexity of the environment, amongst other things. (This has been discussed in the previous sections.) Another goal is to develop a greater understanding of the actual *workings* of trust, so that the concept can be discussed in a precise and meaningful manner. An initial aim was related to the first, in that instilling a knowledge of trust in artificial agents allows the agents to reason sensibly in cooperative (or competitive) situations involving others (trust is predominantly a social phenomenon, despite its undoubted utility in adaptation). Clearly, trust plays a major role in initiating and maintaining cooperative relationships, even if emotions are ignored (as in the example of the wife and the evening meal).[18] This argument is addressed experimentally in chapter 7.

3.4 Part Three — Tools and Research Areas

This part of the chapter presents discussions of the particular aspects of several research areas, from Game Theory to DAI, which may be useful in attempts to experiment with implementations of trust. The first section digresses briefly with a justification of the choice to use mathematics for the formalism, and takes as its justification the relative success George David Birkhoff attained when attempting a scientific understanding and formalisation of the concepts of aesthetics and, perhaps

and less likely to be taken (Rempel *et al.*, 1985; Boon & Holmes, 1991).

[18] In fact, ignoring the emotions involved gains us very little apart from the benefits of not worrying about how emotions work. What trust involves is costs and benefits — the costs to the wife should her husband refuse to cook the tea are great because of the emotions involved. This is easily simulated artificially since we can state that the relationship is close: close relationships bear great expectations, and when great expectations are crushed, the costs are very high, thus trust is decreased by a relatively large amount.

more importantly from the point of view of the present work, ethics. Birkhoff's relatively simple formlułæ show that an "understanding" of such concepts can indeed be arrived at, and that the formalism used can initiate practical and enlightening discussions.

Following this, we present an overview of the particular research approaches that can be used in experimentation concerning a phenomenon such as trust. In no particular order, we see these to be:

1. Traditional methods, such as sociology and psychology.

2. Game Theory and its associated practical implementations.

3. Artificial Life.

4. Distributed Artificial Intelligence.

Whilst this list is not exhaustive, it contains the major areas that are applicable. To be fair, the final three items are inherently inter-disciplinary, taking into account many of the aspects of the first, adding more flexibility as we proceed down the list. We reach the final conclusion that DAI provides social agents, the concept of distribution, distributed intelligence from independent agents, and the concept of interdependence in that the freedom of other agents enables them to affect that of others. With these benefits, we can simulate and test the formalism for trust.

3.4.1 Formalising Aesthetics — George David Birkhoff

The proposed use of mathematics to formalise an agent-centred notion such as trust is not particularly new. The field of aesthetics was examined by the eminent mathematician George David Birkhoff earlier this century.

Birkhoff followed a long tradition in mathematics and philosophy in his attempts to define a formalism for the concept of aesthetics. Indeed, his list of those who had before him contributed to the understanding of the concept is impressive, and contains philosophers such as Pythagoras, Plato and Aristotle, artists such as Michelangelo and Pacioli (who rediscovered the Golden Section), mathematicians, and psychologists (Birkhoff, 1933). The length of the list is an indication of the prevalent thesis that such notions *could* be understood better if mathematics were to be used. Some of those mentioned are artists, particularly musicians (Rameau) and poets (Poe). These are not mentioned because of their art, rather because of their applicability to the scientific study of aesthetics Birkhoff proposes, since "aesthetics, if it is to be scientific, must be approached from the analytic point of view and must concern itself chiefly with the formal aspects of art" (Birkhoff, 1968, page 386).

Birkhoff's work in aesthetics led him to the all-encompassing equation:

$$M = \frac{O}{C} \tag{3.1}$$

Where the concept of aesthetic measure is denoted by M, that of complexity by C, and that of harmony, symmetry, or more generically order, by O. Thus, Birkhoff notes that the general aesthetic experience is to regarded as a compound of three successive phases. The first concerns the preliminary effort of attention, which Birkhoff argues increases with C, the second is the reward, the feeling of value from the experience, M, the third is "a realization that the object is characterised by a certain harmony, symmetry, or *order* (O), more or less concealed, which seems to be necessary for the aesthetic effort" (Birkhoff, 1968, page 321).

Birkhoff's aesthetic equation is remarkably simple, yet beneath it lies a clear understanding of the concept of aesthetics. Not only that, its constituent parts are themselves calculable from their own components. Thus, complexity, C, is based on the physiological-psychological ideas of the time. Birkhoff argues that there is a 'tension' felt when observing objects, and that the tension is greater the more complex is the object. This tension can be measured, or at the least calculated, from partial tensions felt when the perception automatically adjusts itself to view the object. Such adjustments relate to motor nerves, the cerebral cortex, and so forth (*ibid*, page 322). Birkhoff continues:

> ... the feeling of tension always attendant upon perception appears as a summational effect due to the partial tensions which accompany the various automatic adjustments. Thus, if A, B, C... are the adjustments required, with respective indices of tension $a, b, c\ldots$ and if these adjustments A,B,C... take place $r, s, t\ldots$ times respectively, we may consider the sum $ra + sb + tc + \ldots$ to represent the total (negative) tone of tensional feeling... It is this feeling of tension or effort of attention which is the psychological counterpart of what has been referred to as the complexity C of the aesthetic object. In this manner we are led to regard the sum

just written as the measure of complexity, and thus to write

$$C = ra + sb + tc + \cdots$$

Birkhoff, 1968, pages 322–323

Similar discussions are provided for order, O. Thus in fairly straightforward yet inherently sensible stages, Birkhoff summarises the aesthetic experience, and provides several detailed examples of the formalism at work. Of particular interest is his investigation of the aesthetics of rhyme (see Birkhoff, 1933, chapter 8) and that of polygonal forms (see Birkhoff, 1933, chapter 2, also Birkhoff, 1968, pp 334–381, which is a reprint of an earlier paper). Indeed, Birkhoff applied his formula to a great number of aesthetic feelings, and each was well summarised by the formula.

Birkhoff was spurred by his results in aesthetics to attempt to formalise ethics. This is a considerably more difficult task, since ethics, to a far greater degree than aesthetics, are referenced to the morality of the observer. Nevertheless, the psychological assumptions Birkhoff made allowed him to attempt a formalism. The formalism is similar to that for aesthetics. Birkhoff argues that the concept of ethical measure is similar to that of aesthetic measure, and that ethical measure is related to the total good achieved in a situation, thus

$$M = G \tag{3.2}$$

Where M is the ethical measure , the "amount of moral satisfaction based on good accomplished" (Birkhoff, 1968, page 757), and G is the total good achieved. Thus the "ethically-minded person endeavours always to select that one of the possible courses of action which *maximises* the ethical measure G, just as the aesthetically-minded person compares aesthetic objects and prefers those which maximise the aesthetic measure $\frac{O}{C}$" (*ibid*). Birkhoff goes on to compare the concept of aesthetic measure and ethical measure in great detail. This serves to suggest that he envisaged that the two were alike in that they were formalisable, and understandable from relatively simple mathematics. It is this hypothesis that the present work mirrors. That Birkhoff was able to formalise aesthetics in quite a detailed fashion, and to formalise the more complex notion of ethics by a smaller, although significant, amount, is a pointer to the hypothesis that such formalisations can be applied to other aspects of the human or animal psyche, of which trust is one.

3.4.2 Ways of Addressing the Problem

Having discussed the phenomenon of trust, and presented many views on the subject, it becomes clear that a clarification is of use. There are several ways of addressing such a clarification. Some of them have, for other aspects of society or psychology, been attempted before. For example, on various aspects of cooperation and conflict, Game Theoretical approaches, such as Axelrod's experiments with the Prisoners' Dilemma (see especially Axelrod (1984), also Axelrod (1986)). Axelrod's approach involved implementations on computer, as did Dawkins's (1986) example for the Blind Watchmaker, involving human beings artificially selecting attributes which were desirable in computer-bred animals.

The use of computers for simulation and experimentation in such areas as psychology and biology is becoming more widespread. In aspects of biology and more recently sociology (or at least, society), Artificial Life (AL) is a steadily growing research field (Langton, 1990a) with applications in many areas. Some aspects of AL, such as Genetic Algorithms (GAs), first mooted in Holland's seminal work on the subject (Holland, 1975) and developed widely since are becoming more and more successful in their own right, and have found applications in areas as diverse as Neural Networks and the Prisoners' Dilemma game (Axelrod, 1987). AL is also shedding light on societies of simple animals, such as ants (Drogoul & Ferber, 1992), whilst providing mechanisms allowing greater dissemination of information across distributed databases (Witten & Thimbleby, 1990; Witten *et al.*, 1990). In short, Artificial Life is moving forwards all the time.

A spinoff, it could be argued, from the social side of Artificial Life, whilst inheriting all of the benefits of intelligent agency from Artificial Intelligence, is Distributed Artificial Intelligence (DAI). DAI benefits from many productive years of research in AI, whilst adapting it's premises to take into account the idea of a distributed system of artificial intelligent agents. In a sense, then, it is AL with intelligence. Indeed, often the distinctions blur, both between DAI and AI, and between DAI and AL. Whatever the case, much of the power of DAI has yet to be exploited.

3.4.3 Traditional Research Methods

Traditionally, computers have been on the periphery of the social sciences, used for data collection and assimilation, rarely for simulation. More and more, however, computers are becoming part of the repertoire of research tools proper of the social scientist. Political philosophers and psychologists, for example, benefit from computer implementations of Game Theoretic approaches such as the Prisoners' Dilemma (Axelrod, 1984; Axelrod, 1987). The fact remains, however, that this usage is still on the periphery, and the methods used are for the most part not well integrated, both inter- and intra-subject. In other words, philosophy, sociology, psychology, and other social sciences are lacking in integration, with the consequent loss of a clear direction.

This lack of integration is a clear hindrance to the furtherance of understanding of social concepts such as trust. The study of such phenomena is inherently broad and does not lie within the boundaries of any specific research area. As an example, Aritificial Intelligence, or its study, requires a broad repertoire of understanding, from philosophical aspects to the psychology of mind. Phenomena such as trust and morality lie in as broad a spectrum. The result is that in order to study such phenomena accurately, a much broader approach is necessary than any one research area can provide. Distributed AI, as discussed above, provides such an approach, incorporating aspects of sociology, psychology, AI and philosophy, amongst others (Bond & Gasser, 1988).

3.4.4 Game Theory

Game Theory is an important sphere of research in psychology, politics, and many other social spheres. It has, in the past, provided many important insights into behaviour in *zero-sum* games — games of pure conflict (Schelling, 1960). Indeed, in the sphere of non-zero-sum games also, Game Theory has had much to offer.

Game theoretic approaches, however, do not explicitly consider the wider aspects of societal existence which are of importance to concepts such as trust, which in itself would not exist without society. In addition, many Game Theoretic approaches are inherently confrontational: the Prisoners' Dilemma, for example, assumes two 'enemies' facing each other, with no prospect of real communication and no real method of encouraging cooperation for any single interaction. The iterated PD goes some way to include a method for encouraging cooperation in the form of the 'Shadow of the Future' (Axelrod, 1984), but is still inherently confrontational. This being so, it is often the case that the agents who do best are those who are perhaps less moral and defect first (Behr, 1981). Trust is not a confrontational phenomenon. Indeed, it serves in many cases to avert confrontations, and provides agents with the means to judge others favourably and put themselves in others' hands with some confidence[19] as to the outcome of that decision.

A Game Theoretic approach to the study of trust does provide some answers — some such studies have been entered into, and Game Theoretic notions of some of the workings of trust do exist (see Boon and Holmes (1991) for an example). They are limited, however, and rely often on vague language constructs in attempts to clarify many of the notions involved. Game theory clearly does not provide sufficient tools for the understanding of trust, although it does go some way towards understanding the workings of interactions between individual agents, and this is of some use.

3.4.5 Artificial Life

Artificial Life is a strong, growing research field with applicability in all of the above discussed areas. In Game Theory, we can benefit from sophisticated implementations of theoretical puzzles. In biology, Artificial Life may help us ultimately understand much of what is still hidden by showing us new forms of life (Langton, 1990b). In anthropology, behaviours can be studied and experimented upon, producing, once again, a greater understanding of the world and its inhabitants (see, for example, Bill Coderre's PetWorld (Coderre, 1989), or Pattie Maes's experiments with the selection of actions (Maes, 1990c; Maes, 1990b; Maes, 1991; Maes, 1991), see especially Maes (1990a). Also on action selection, see Tyrell (1993)).

Artificial Life's major drawback concerns its lack of intelligence, the lack of, in this instance, purposive deliberation on the part of an agent. Indeed, AL's main strength is in showing the arising of emergent properties — properties of a system as a whole derived from, but not necessarily present in, the properties of the agents which constitute the system (Forrest, 1990). By definition, emergent properties can never be deliberative (although they can be made use of when they arise, and when a 'use' is detected for them (Wavish, 1991)). Thus, intelligent, or at least purposive, deliberations are outwith the scope of strict AL.[20] It is widely accepted that, without

[19]The word 'confidence' has been used earlier, and criticised by Luhmann, who states that it does not imply consideration. This is not that form of confidence, rather it is confidence in one's judgement that the other is trustworthy.

[20]Whilst emergent behaviour can arise in DAI, or any other complex system. Also, DAI, as a superset of AL, can make use of reactive aspects of agents (Chapman & Agre, 1987; Spector &

anthropomorphising, trust is such a deliberation.[21] The answer, then, appears to be the use of DAI for a study of trust.

3.5 Distributed Artificial Intelligence

DAI is a research field which is growing in applicability and relevance, both to Artificial Intelligence and to computing as a whole. In short, it involves concepts of distribution, intelligence, society, independence, graceful degradation, and localised decision making. It is, then, an ideal tool for the study of trust. The following section summarises the strengths of DAI as a tool for the study of an implementation of trusting behaviour in artificial agents, taking into account the criticisms of other approaches discussed above.

3.5.1 The use of DAI as a Research Tool

With independent, rational intelligent agents, we are presented with the ideal tool for researching implementations of trust. We summarise the reasons here:

- Agents are assumed to be (pseudo-) intelligent. Since in societies of agents, intelligent behaviour must include considerations of the beliefs and attitudes of others who may affect the agent who is considering, trust presents an agent with an extra capability in that direction. In addition, without seeking to anthropomorphise, trust as an affector of behaviour in artificial agents may present interesting viewpoints on human agents.

- Agents are assumed to be (pseudo-) rational. Whether or not trust is rational depends on many things. In Deutsch's investigations, the story of the Princess and the Tiger presents an interesting view of this, with some of the forms of trust (e.g., trust as despair, trust as faith) not necessarily being rational, and others (trust as confidence) being rational in many cases. For rational agents, a

Hendler, 1990; Downs & Reichgelt, 1990)

[21] It is granted that, in animals, trusting behaviour, such as in the form of delayed reciprocation, can and does take place (Harcourt, 1991). Indeed, in this work, we take such an open view of trust. The fact remains, however, that the human race makes most efficient use of the phenomenon, to the extent that it has become a part of our everyday existence, and we deliberate about it, both consciously and unconsciously.

goal would be the implementation of rational trust. Trust as confidence supplies this viewpoint. However, instances of irrationality in this area of trust can be studied more closely in supposedly rational agents.

- Agents are generally assumed to be cooperative (von Martial, 1992). In fact, agents are generally assumed to be trustworthy also, and clearly neither assumption can be justified in the 'wider world' that exists outside the research labs. Nevertheless, there are situations within which cooperation is necessary (Connah & Wavish, 1990; Marsh, 1992), and others where, although cooperation may be necessary, there is a choice about who to cooperate with (Marsh, 1992). In both of these types of situation, the behaviour of the formalism can be tested and refined.

- Agents are distributed. Trust is a societal concept (Luhmann, 1979; Baier, 1986), and distributed agents, with non random interactions between them (either through choice or direction) are a society in themselves (Bond & Gasser, 1988; Gasser, 1991). Thus trust is not out of place and can be observed in interactions within such artificial societies.

- Agents are generally independent. Trust is a means of coping with the freedom of others (Gambetta, 1990a; Luhmann, 1979). Within a society of independent agents, its behaviour can be readily investigated and observed, and anomalies detected.

- There is a wide range of potential and actual applications for DAI, from Air Traffic Control (Cammarata *et al.*, 1983) to Open Informations Systems (Hewitt, 1991). Trust can potentially be implemented and observed in realisations of many such spheres, and its behaviour and influence detected and refined over time.

3.5.2 Amalgamation

The implementations to be discussed in chapter 7 are not strictly DAI. In fact, they are very closely allied to Game Theory in that they use the Prisoners' Dilemma situation, but are an amalgamation with many of the concepts of DAI in that the agents are independent, social, and geographically distributed, with some control of movement.

Thus, the agents are limited in terms of intelligence, although they are knowledgeable of the aspects of the Prisoners' Dilemma, particularly in terms of payoffs and utility, and so they are more complex than the usual entities of Artificial Life. Future work (see chapter 8) envisages more detailed intelligent trusting agents.

3.6 Summary

Trust is an issue which is close to all members of society, however it is used (Luhmann, 1979). It is thus an extremely complex area, and one which is not properly understood or researched (Luhmann, 1990). This chapter has presented a summary of many of the prevailing views of trust, from fields such as sociology, social psychology, and economics, to name but some. It is inevitable, with trust being so all-pervading, that such a review cannot cover the entire sphere in detail, and will obscure some aspects of that field. What was attempted then was to present the view of trust that several more prominent researchers in the field held, comparing and contrasting these views. The chapter then discussed aspects of trust which, although considered of importance, were not easily classified in these terms. The result is a comprehensive, although not complete, review of the work done concerning trust, and the thoughts of several prominent minds. A complete review is both outside the scope of the present work and unnecessary in the attempt to provide an understanding of the concept.

The present work is concerned with formalising the concept in a way that enables its inclusion in artificial agents. Several methods for examining such an idea were critically discussed, with the conclusion that the field of DAI presented unique opportunities in the form of pseudo-rational, intelligent, sometimes cooperative, independent agents. Such agents, it is argued, will also benefit from the inclusion of a concept of trust in their decision-making repertoires.

To that end, the following chapters present and work with the formalism that has been developed in this work. Chapter 4 presents the formalism, based on an amalgamation of much of what has been presented here, and refined much over the past three years. Chapter 5 discusses the formalism, showing examples of it at work. Chapter 6 uses the formalism in a descriptive way, discussing some principles for trust. Chapter 7 culminates with discussions of implementations of the formalism in simple artificial agents. Such implementations not only follow the correct experimental path for work of this kind, but also justify the claim that the formalism can be implemented and produce agents whose trusting behaviour is as can be expected.

Chapter 4

An Example Heuristic Formalism

"Justice is represented by a square number"

Pythagoras.

"...values are not the least vague when you're dealing with them in terms of actual experience."

Pirsig, 1991, page 63.

4.1 Discussion

From the previous chapters, it is clear that a unified theory for trust is lacking. In order to develop such a theory, and associated principles, it is necessary to provide a precise means of discussion about the phenomenon. Deutsch (1973, developed further by Golembiewski and McConkie (1975)) can be seen as starting to develop such a means of discussion in the various aspects of trust that he brought together into one solid consideration of the subject. The present work provides a formalisation[1] of trust with which to continue the discussion in a precise, unambiguous manner. As was suggested in chapter 2, the introduction of a new formalism has potential problems, not least that the formalism will be seen by many as too restrictive, particularly with something as rooted in our subconscious and unconscious thoughts as trust. Acknowledging those problems, we argue that the formalism as it stands can help in discussions about them, leading to better versions of itself.

[1]There is perhaps confusion between the use of the words 'formalisation' and 'formalism'. For the remainder of the thesis, we discuss a 'formalisation of trust' in terms of a 'formalism for trust'. The two are considered to be interchangable here.

This chapter introduces the formalism that has been developed. The early development of the formalism was largely based on discussions of trust in the literature (see chapter 3). However, the later work built upon the initial formalism using experimental results and other observations, including discussions with others whilst applying it to areas other than cooperation in DAI (e.g., as in Thimbleby *et al.*, 1993). What follows, then, is a formalism that is considerably more applicable than when it started out (and as presented in Marsh (1992), Marsh and Thimbleby (1992)) without losing any of the intuitively appealing aspects of trust that it first described, such as the three forms of trust (see below). The formalism consequently has more applicability to the consideration of the phenomenon of trust in social situations than it first achieved.

The main consideration in the introduction of such a formalism is which approach to take. In this work, the aim was to provide a formalism which was as simple to understand as possible, whilst preserving its expressive power (the use of Occam's Razor). One reason for this was to allow an implementation based very closely on the formalism to be developed. This consideration led to the decision to use simple probabilistic methods for describing trust. This argument is put in chapter 2. A differing approach to the concept of trust — that of harmony, or *Wa* — is discussed in chapter 8.

4.1.1 Overview of the Chapter

The formalism, as might be expected of something describing trust, is large in the sense that extensions are possible, have been described, and continue to be made. This chapter presents the formalism 'as is,' in the hope that it will spur others to extend it and correct it further, and the knowledge that such work is indeed possible. To allow a greater understanding of the formalism and associated formulæ, we present it in a stepwise fashion. Thus the next section will present the initial aspects of the consideration, describing how we represent trust at all in a formalism. Following that, we present the major aspect of the formalism — its application to agents deliberating in a potentially cooperative situation. Acknowledging the restrictions in the formulæ, we proceed to extend them, attaching other considerations, finally achieving a solid formulation which considers temporal constraints, differing situations and the similarities between them, and the past behaviour of other agents and the environment as a whole (although such a consideration is, in this chapter, implicit). Chapter 5 extends the formalism further, showing its use for describing considerations of trust, and chapter 6 presents some principles which trust in general adheres to.

4.2 Initial Considerations

In this section, we present the basic notation used in the formalism.

4.2.1 Agents and Situations

We represent agents by $a, b, c, \ldots$ Individual agents are members of the set of all agents, $\mathcal{A}$. Particular subsets of that set can be represented with $\mathcal{B}, \mathcal{C} \ldots$ In particular, we can represent *societies*, or *communities*, of agents. We define a society of agents as a number of agents (greater than 1) which is grouped together according to some metric. This is in keeping with the definition found in van den Berghe (1980) which defines society as "a group of conspecifics bounded by a zone of much less frequent interactions than the rate which prevails between members." (page 77), with the notion of interactions being replaced by that of some measurable metric. In fact, interactions provide us with the ideal measure of such a notion of society, and allow the idea of 'nested' societies. Thus, we can be members of several societies, each of which may be either independent or members of others. For example, I am a member of the society of the village in which I live, but also of the country at large (the 'normal' view of society). In addition, I am a member of a more global community, or society, which communicates via the Internet. Thus, societies are no longer bounded by physical space. The same applies to agents. For the purpose of the formalism, we represent societies with $\mathcal{S}_1, \mathcal{S}_2 \ldots \subset \mathcal{A}$.

Agents find themselves in particular *situations* by definition. This is because we define a situation as a point in time relative to a specific agent. Different agents in the 'same' situation will not consider it from the same point of view. Halpern and Moses (1990) illustrate this point with the story of children playing, some of whom have mud on their foreheads, and some who don't. For each child in the situation, it is different (each can see every other child except themselves). Thus, in a particular situation, agents may have only incomplete knowledge about that situation, yet some knowledge may be common (there are some children with mud on their foreheads — which ones is more difficult to answer, since we don't know if we're one of them). Since situations are agent-centered, or subjective, we represent them with Greek letters, with subscripts for the agent concerned: α_x is situation α from x's point of view, β_y is situation β from y's. In the formulæ that follow, the subscript is often dropped. This is because it is evident which agent is doing the considering. Situations are taken to be members of the set of all possible situations in the world (a very large and open ended set!) notated S.

4.2.2 Knowledge

It is important to be able to reason about whether one agent 'knows' another. In other words, whether the two have met at some time. It is important for the formalism because with such an explicit knowledge, we can make certain claims about trust (see chapter 6 for a discussion of knowledge in trust). As far as agents are concerned, the concept of knowledge is important in the sense that considerations of trust imply considerations of knowledge of the trustee, at least after the first interaction between two agents. That one agent (x) knows another (y) is given by $K_x(y)$ — they have met at some time, and x can remember it.[2] One point worth making here is that, whilst the items in the formalism have explicit values, some need not be considered. Knowledge in this sense is a boolean concept — one either knows someone or one does not know them. Thus, K has a value of 0 or 1, or true or false. The notation allows us to ignore this, and just write $K_x(y)$ for x *does* know y, and $\neg K_x(y)$ for the opposite. This does not preclude the use of more specific values (representing partial knowledge, for example) where necessary.

4.3 Trust

We have separated trust into three different aspects: basic, general, and situational trust.

4.3.1 Basic Trust

Agents are considered to be trusting entities. As such, they will have a 'basic' trust (Boon & Holmes, 1991; Govier, 1992). This is derived from past experience in all

[2] For a discussion of memory in trusting agents, see later in this chapter.

situations, indeed through the agent's entire life of experiences. It is represented by T_x for the basic trust of agent x. It has a value in the range $[-1, +1)$, thus $-1 \leq T_x < +1$. It is *not* the amount of trust an agent has in any other agent, or situation, or the environment; it is simply representative of the general trusting *disposition* (Boon & Holmes, 1991) of the agent. The higher it is, the more trusting is the agent. Considering agents as trusting entities, this allows us to simulate a basic 'disposition' to trust someone or something that has only just been encountered. If we consider the agent to be adaptive in the sense that it learns from past experience, we would then expect that disposition to be dependent on what has happened to the agent in the past. Good experiences lead to a greater disposition to trust, and vice versa (Boon & Holmes, 1991).

There are possible exceptions to this, however. Pathological forms of trusting can result in the behaviour of trust not being so clear cut. In Marsh (1994), we present a view of some dispositions to trust, among which are optimists, pessimists, and realists. We argue that optimists are general trusters whose disposition to trust is relatively inflexible in a downward direction, despite past experience. Thus, their trust in others can only increase, never decrease. The opposite is true for pessimists. We continue with this argument later in the chapter.

4.3.2 General Trust — Trust in Agents

Given two agents, $x, y \in \mathcal{A}$, to notate 'x trusts y,' we use: $T_x(y)$. Given that we are using a value-laden aspect of trust, this also has a value in the interval $[-1, +1)$. Thus $-1 \leq T_x(y) < +1$. The value represents the amount of trust x has in y here. It is not relative to any *specific* situation (see below), it simply represents general trust in another agent. A value of 0 means x has no trust in y (indeed, may not actually know y at all, although the existence of a representation for $T_x(y)$ would seem to imply this knowledge.). A value of -1 would represent a negative trust — complete distrust. Informally, the value represents the probability that x will behave 'as if' he trusts y. In other words, x expects that y will behave according to x's best interests, and will not attempt to harm x (see Thimbleby *et al* (1994), Marsh (1994), Boon and Holmes (1991)).

No Trust and Distrust

The two aspects, of zero trust (or no trust) and distrust, are not the same. The reasons for a situation in which one agent has a zero trust in another may be several:

- The trusting agent may not know the trusted agent.
- The trusting agent may be impartial with respect to the trusted agent.
- The two may have just met, with the trusting agent having $T_x = 0$, thus assigning 0 to general trust.
- Experience may have led to the decision to allocate a trust of 0 in the other. This may be because the other was previously trusted in a positive fashion and has behaved badly, or *vice versa.*

This list is not exhaustive, and serves to illustrate various paths to the judgement that an agent is 'not' trusted (trust value 0). It also serves to illustrate that, in some circumstances, the value can be allocated because of incomplete knowledge of the other agent. However, complete distrust is arrived at through some judgement at least. It is an extreme statement to make about someone to say that you have complete distrust in them, and is usually arrived at through careful consideration of past events with that person. In particular, you most certainly do *know* that person. This seems to suggest that a value of -1 should not be ascribed to any agent on a first meeting. Whilst it is possible, given that the basic trust value of an agent may be -1, it is hoped that the futility of such an ascription of distrust is clear — the agent who made such a 'rash' judgement would have severe problems getting anything collaborative done!

Blind Trust

It can be seen that whilst a value of -1 for complete distrust is possible, a value of $+1$ for complete trust is not. There is a justification for this which is founded on philosophical ideas of trust. Trust implies a consideration of something or someone (Broadie, 1991). Even a value of 0 ascribed to something (when the agent *knows* of that something — see also chapter 6), implies some consideration of that something in order to arrive at the decision that, perhaps, more consideration is necessary. Either that, or it acknowledges that such consideration is not possible, and reserves judgement.

However, in this system, a value of $+1$ means absolute trust — blind trust. It is in the name blind trust that the problem becomes apparent. Since trust necessarily means that we have searched for evidence to believe (in) something (Broadie, 1991), blind trust implies blind acceptance, a sheep-like acceptance of what is. Thus, it also means that the thing being trusted is not considered at all (why bother, if one trusts it blindly?). Blind trust is not trust, as it does not involve thought and consideration of things (this definition is similar to Luhmann's notion of confidence (Luhmann, 1990)). A value of $+1$ is thus not accepted here. Nor is it generally acceptable in everyday life, where credulity is seen to be problematic (Dasgupta, 1990).

The value for complete distrust, -1, is also questionable, from another point of view. When a value of -1 is reached, we are implying that no-one could be trusted less from our point of view. Just as no-one is perfect, so no-one should be completely imperfect. Although this seems an unnecessarily optimistic view of the nature of things, it is not the basic reason why a trust value of -1 is possibly unsound. Giving an absolute 'perfect' distrust means that no-one can be trusted less. It is unwise to make such a judgement, since we do not have knowledge of everyone and everything in the world (Luhmann, 1979). Round the corner, it is possible, is someone even less trustworthy than this person. These considerations apply equally to blind trust, of course. The reason complete distrust is present in the formalism is that it is not contrary to the idea of trust that we have, specifically concerning consideration. Complete distrust, or incredulity, almost certainly does require consideration before and after it is ascribed, and continual consideration should be given to the matter. Blind trust implies no consideration, thus it is not included in the formalism.

4.3.3 Situational Trust in Agents

As was mentioned above, agents are based in situations. A situation means something different to each agent experiencing it. In addition, it is the case that different situations will require different considerations with regard to trust, and most will come out with different values for trust, even in the same person; whilst I may trust my brother to drive me to the airport, I most certainly would not trust him to fly the plane! This suggests that agents should consider others relative to the situation they find themselves in (indeed, to the things they envisage the situation to involve). Thus, we have a representation for the amount of trust an agent has in another in a given situation.

This, and the previous value, are arrived at through consideration (see below). The notation for 'x trusts y in situation α' is: $T_x(y, \alpha)$. Here, we drop the x from α_x, since the whole value is clearly taken from x's point of view. Once again, this takes a value in the interval $[-1, +1)$. Rempel and Homes (1986) state that trust is in people, not their actions. The view of situational trust extends that view of trust, stating explicitly that trust is in people *in specific situations.* We still have trust in people rather than their actions, but we add power to the definition by allowing considerations over time in differing circumstances. Situational trust is discussed in more detail later in this chapter.

4.3.4 Importance and Utility

The rational economic actor attempts to maximise utility (Simon (1955), see also Preston (1961), Sycara (1988)). In our formalism, this also holds, although considerations of utility will be allied to considerations involving trust.

The notation for utility is similar to that for knowledge, thus, for the amount of utility x gains from situation α, we write $U_x(\alpha)$. $U_x(\alpha)$ has values over the interval $[-1, +1]$. The values can be normalised over this range. We take utility to be based in expected utility theory, with for example the overall utility of a particular situation, over all outcomes, as being the mean of these utilities of the outcomes (Zeckhauser and Viscusi, 1990). This implies that the agent concerned should consider every possible outcome of a situation. Whilst this may well be feasible in situations with few outcomes that are known with some certainty, there will be situations where some outcomes are not known, or with a probability of occurrence which is not known (*ibid*). An agent can thus rely on, for example, a weighing up of the costs and benefits that it estimates the situation holds. For a learning agent that adapted according to the results that it obtained from situations, it should not be too long before these estimates became fairly accurate.

Related to utility is the concept of the importance of a situation. On consideration, it may seem that the importance and the utility are one and the same. This is not the case, although the distinction is not obvious. In particular, utility is generally measurable, or at least relatively straightforward to find an estimate for, whereas importance is an agent-centered or subjective judgement of a situation on the part of the agent concerned. For example,[3] the utility of winning at dice in Las Vegas is potentially enormous, whereas the importance is negligible, since it is not rational to expect to win. Were the odds stacked in your favour (loaded dice, perhaps?) then it becomes vastly more important that you play, whilst the utility of winning may well stay the same. In addition, whilst the concept of utility is an accepted measure of the actual outcome of a situation, importance is a subjective measure of the expected benefits to be gained from a situation under consideration. The importance of two identical events on different days may well be assessed differently for the same agent. Importance allows us to represent the fact that things outwith the situation change in the world (Spector and Hendler, 1990). The agent itself may change, it may receive specific orders to carry out some action which make that action much more important now than it was yesterday, and so forth. In addition, trust cannot be based on rationality alone (Herzberg, 1988; Lagenspetz, 1992). The subjective concept of importance allows something additional to rationality to be considered. Importance gives the formalism added prescriptive and descriptive power.

We represent the importance of a situation α for agent x with $I_x(\alpha)$. It has a value over the interval $[0, +1]$. Negative importance is not considered here. This is because for the purposes of this work, we take situations of negative importance to be the opposites of situations of positive importance — thus "it is important that this is *not* done." Further complexity is outwith the scope of this work.

To summarise the notation in an accessible manner, we present it in table 4.1.

4.4 Temporal Considerations

The above notation is adequate for discussing trust in a static situation, where memory is assumed to be perfect, and trust alterations have been carried out 'elsewhere.' Time matters in the real world, however. Indeed, trust is needed *because* transactions between agents take place over time (Coleman, 1990). As it stands, there is no means of representing time in our formalism, We introduce a temporal index to allow us to refer to, for example, how much x trusts y now, compared to how much that was last week, or some time ago.

Providing a temporal index is a relatively simple task. The reason why the for-

[3] A similar example to this was suggested in a discussion with Harold Thimbleby.

Description	Representation	Value Range
Situations	$\alpha, \beta, \ldots$	
Agents	$a, b, c, \ldots$	
Set of agents	$\mathcal{A}$	
Societies of agents	$\mathcal{S}_1, \mathcal{S}_2 \ldots$	
	$\mathcal{S}_n \in \mathcal{A}$	
Knowledge (e.g., x knows y)	$K_x(y)$	True/False
Importance (e.g., of α to x)	$I_x(\alpha)$	$[0, +1]$
Utility (e.g., of α to x)	$U_x(\alpha)$	$[-1, +1]$
Basic Trust (e.g., of x)	T_x	$[-1, +1)$
General Trust (e.g., of x in y)	$T_x(y)$	$[-1, +1)$
Situational Trust (e.g., of x in y for α)	$T_x(y, \alpha)$	$[-1, +1)$

Table 4.1: Summary of the basic (non-temporal) notation.

malism for static situations does not possess the temporal index is that we consider it to be necessary to be able to reason without the index, since the index adds a dimension that, in some cases, we do not wish to exist. For example, when implementing a computer simulation based on the formalism, the temporal constraints need not be there, since they are already present in an ongoing simulation. The computer provides the necessary fluidity in terms of time to allow us to forget consideration of temporal aspects in the formalism. Defining the formalism without the temporal index makes it clear that we don't need it to make the formalism work. It is an addition, extra descriptive power is given that in some cases is not needed.

The addition of the temporal index is achieved by the addition of a superscript to the items we wish to index which represents a specific moment in time. Thus:

$$T_x(y)^t$$

Represents how much x trusted y at time t. Table 4.2 gives the time-extended version of the notation. As can be seen, it differs very little from the basic notation.

Description	Representation	Value Range
Situations	$\alpha, \beta, \ldots$	
Agents	$a, b, c, \ldots$	
Set of agents	$\mathcal{A}$	
Societies of agents	$\mathcal{S}_1, \mathcal{S}_2 \ldots$	
	$\mathcal{S}_x \in \mathcal{A}$	
Knowledge (e.g., x knows y at time t)	$K_x(y)^t$	True/False
Importance (e.g., of α to x at t)	$I_x(\alpha)^t$	$[0, +1]$
Utility (e.g., of α to x at t)	$U_x(\alpha)^t$	$[-1, +1]$
Basic Trust (e.g., of x at t)	T_x^t	$[-1, +1)$
General Trust (e.g., of x in y at t)	$T_x(y)^t$	$[-1, +1)$
Situational Trust (e.g., of x in y for α at t)	$T_x(y, \alpha)^t$	$[-1, +1)$

Table 4.2: Summary of the temporally-indexed notation.

4.5 Using the Notation

Having introduced the notation, we can develop it to allow useful discussions about trust, in addition to giving an agent the means to reason with and about trust in inherently cooperative situations. We focus on cooperation in this chapter since DAI is itself inherently cooperative (see chapter 1). This is natural, since we want our agents to be cooperating to get things done. If they didn't, we would be as well assuming the agents were monolithic AI systems, with no interaction with each other at all, since none would be needed. Not only is this unrealistic, but even for traditional AI systems (Hewitt, 1992), it is becoming less applicable (Gasser, 1991).

In considering cooperation there are some questions which need to be asked (Marsh, 1992):

- With whom to cooperate.
- To what extent cooperation should be extended.
- When to cooperate.

The basic question is, given the choice between cooperation and non-cooperation, whether to cooperate with a specific agent or not. In the extensions below, we provide a means of arriving at an answer to that question in a situation (α) in which one agent (x) is considering whether or not to cooperate with another (y). The basic assumptions we make are:

1. That x has a choice — he can say no if he wishes (for a consideration without this assumption, see the following chapter, section 5.2.2).
2. That there is someone (e.g., y) to cooperate with in this situation.
3. That x has a knowledge of y ($K_x(y)$ is true).
4. That x is not 'in debt' to y. This rules out any consideration that x helps to reciprocate for something y has done in the past. See page 100.
5. That x has a knowledge of the situation, in other words, that x perceives similarities between this situation and others he has experienced.[4]

4.5.1 Determining Situational Trust

Of the three types of trust: basic, general and situational, situational trust is of most importance when considering trust in *cooperative* situations. Indeed, the basic assumption is that, if the situational trust is above a certain threshold (see section 4.7) cooperation will ensue.

In order to estimate situational trust, x will need to consider several aspects of the situation. It is a rational thing to do to try to increase utility as far as possible (Simon, 1955; Sycara, 1988). Whilst this is an accepted definition of economic rationality, it is felt that it lacks elements in decision making which should be present, specifically agent-subjective measures about the situation. It is for this reason that we add a consideration of the importance of the situation to that agent. A situation with large utility but little importance would, in this scheme, be less likely to be undertaken than one with utility large and a high importance. Conversely, a situation with a high importance and a low utility has no more chance of acceptance than the high utility low importance situation.

Since our agents are basically trusting entities, we need to consider trust in our formula. The reasoning is that in order to estimate situational trust in an agent

[4]Although this may not be of interest in this particular working, it becomes more important as we extend the formalism further.

(the trustee), the truster has to use knowledge of that agent in other settings, other situations, and so forth. This information is embedded in the general trust values the truster has of the trustee. These are the $T_x(y)$ values. The estimate of general trust is notated $\widehat{T_x(y)}$ for the amount x trusts y. It is x's estimate after taking into account all possible relevant data with respect to $T_x(y, \alpha)$ values in the past; i.e., if t is the present time, x will take knowledge from all situations $T_x(y, \gamma)^T$, with $T < t$, and γ similar or identical to the present situation α (see Dechter (1984) for a consideration of similarities between situations). If we consider a series of these values, there are several methods of choosing how to work out which estimate to use. These will be discussed further in section 4.6.

We come now to the formula for estimating situational trust in another agent. In this example, we take x to be considering y, having a knowledge of y, and carrying out the consideration with respect to situation α. There are no temporal considerations in this example. To estimate situational trust, x uses:

$$T_x(y, \alpha) = U_x(\alpha) \times I_x(\alpha) \times \widehat{T_x(y)} \quad (4.1)$$

Informally, we define the trust of an agent x in another, y in some given meeting as the probability weighted by UI that x acts to achieve an outcome as if it trusts y (Thimbleby *et al*, 1994).

Analysing the Formula

The formula for situational trust highlights some problems which formalising trust may hold. It is clear that negativity poses problems here, since the multiplication of two negatives results in a positive value. For this formula, table 4.3 presents a detailed examination of final values of the situational trust variable for different extremes of its components. Since this is an important aspect of formalisation, certain formulæthroughout the thesis will be flagged for discussion, and general concepts are further addressed in section 9.6. In addition, the problem of negativity is partially addressed in section 8.6.1.

		$U_x(\alpha)$				
		-1	-0.5	0	$+0.5$	$+1$
	-1	$+1^{1}$	$+0.5^{2}$	0^{3}	-0.5^{4}	-1^{5}
	-0.5	$+0.5^{6}$	$+0.25^{7}$	0^{3}	-0.25^{8}	-0.5^{9}
$T_x(y)$	0	0^{3}	0^{3}	0^{3}	0^{3}	0^{3}
	$+0.5$	-0.5^{4}	-0.25^{8}	0^{3}	$+0.25^{10}$	$+0.5^{9}$
	$+1$	-1^{5}	-0.5^{9}	0^{3}	$+0.5^{9}$	$+1^{11}$

Notes:

[1] Represents one way of behaving, which is machievellian. There may be others, for example, deciding not to cooperate at all if any of the two determinants is negative.

[2] This is machievellian again, mediated by the size of $U_x(\alpha)$.

[3] Again, this is one way of looking at the problem, and results from the chosen operator (multiplication). It seems sensible enough, however — a situation of zero utility or zero trust may well result in indifference as far as situational trust is concerned, no matter what the value of the other determinants. There may be other ways of addressing this situation: this is an avenue of further work.

[4] This is sensible behaviour, mediated by the multiplication operation.

[5] Again, this is sensible, although there may be other ways of attaining a sensible value for differing dispositions.

[6] This is another example of machievellian behaviour.

[7] This is as 6, but see note 10.

[8] This is an example of sensible behaviour for trust.

[9] Again, this is sensible behaviour.

[10] This is an oddity which arises from the use of multiplication and fractions. The result is lower than each of the values for utility and general trust, and will be further reduced by the importance. Other operators (for example logarithms) would perhaps remove this oddity.

[11] This is sensible behaviour for trust as we see it.

Table 4.3: Examination of the formula for determining Situational Trust.

4.5.2 Problems with Importance

As formula 4.1 stands, the more important the situation, the more the final value for situational trust. Thus, we take the view that important situations have to get done (see section 4.7). Paradoxically, this means that the less important a situation is, the lower the situational trust will be, leading to a situational trust of 0 for an importance of 0. The method used for determining the cooperation threshold (section 4.7) takes this into account however, leading to a lower cooperation threshold for low importance. The problem of an importance of 0 is to some extent mitigated by the formula for the cooperation threshold, which will also equate to 0 if the importance is 0. The agent, then, will always cooperate if the importance is set to 0. This is not necessarily such a good thing, however. We leave this problem for future work. In the formulæ below, we assume that this consideration has been undertaken by the agent concerned, thus the value is acceptable. The same considerations should be taken with the cooperation threshold. For more on this, see section 4.7.

4.5.3 Extension to Situational Trust Formula

Trustworthy people are more likely to trust (Rotter, 1971; Deutsch, 1962). The converse holds also: untrustworthy people will generally trust less. Whilst these results have been obtained by experimentation, it would seem intuitively natural that, if you were not capable of being trustworthy, you would expect others to be the same, thus you would not put yourself into a situation where you were in the hands of others. It follows, then, that if we as a trusting agent can somehow estimate how much we are trusted by those we are considering, we can come up with a fair idea of whether they are trustworthy or not. At the very least, we would realise we weren't quite as trusted as we first assumed! In people, it has been found that cooperation (hence trust) is "low when situational factors conspired to make learning about the other person difficult." (Boon & Holmes, 1991, page 196). It's a question of determining the other's motives. Trust is just one of these, but in a cooperative situation, it is of paramount importance (Deutsch, 1962).

The problem is finding out how much we are trusted. Short of asking, there is no way of knowing, and even should we ask, we are not guaranteed the correct answer, particularly if the trust is low. Whilst taking into account that this is an important problem, there is, then, little we can do about it at present.[5] After all, if agents were transparent enough for us to know how much they trusted us, then they would be so transparent that there would be little need for trust in the first place! This being the case, we stick with the workings of formula 4.1. For completeness, however, we present here the possible extension to those formulæ which would take into account an agent's (x here) estimate of how much it is trusted by another (y). This is notated $(\widehat{T_y(x)})^x$:

$$T_x(y,\alpha) = ((U_x(\alpha) + \widehat{T}_x(y)) \times (I_x(\alpha)) \times (\widehat{T_y(x)})^x$$

There is a further problem here. If x's estimate of how much y trusts him is useful, then by definition so is x's estimate of y's estimate of how much x trusts y: $(\widehat{\widehat{T_x(y)}})^{y^x}$! This recursion is potentially infinite, but the amount of recursion is limited since each additional recursion produces data of less and less importance in the scheme of things. Though in the limit, the 'fixed point' might be completely different from the original estimate, it is not obvious what this would mean in practice, since no agent could compute it in finite time.

4.6 Agent Dispositions

In humans, there are differing 'orientations' (Boon & Holmes, 1991) suggesting what kind of things they might do in a given situation. This section discusses orientations in terms of how an agent estimates $\widehat{T_x(y)}$ in formula 4.1. This estimate was presented as x's estimate after considering all possible values for $T_x(y)$ in the past. Little was said about the different means of obtaining this value. We present some here, along with an examination of the orientation or disposition of x that they may signify.[6] We focus on three major statistical methods of obtaining the value. They are, the mean, the maximum and the minimum. The discussions below are taken from Marsh (1994a).

[5]The problem is inherent in trusting relationships, and is not limited to the formalism. Indeed, the formalism may help artificial agents find a way around the problem by allowing them to disseminate trust readily.

[6]Disposition and orientation are used interchangeably here. Thus, the orientation an agent has refers to his disposition towards doing something.

4.6.1 Maximum Estimate — Optimism

The optimist expects the best in all things (Marsh, 1994a). He looks for the best in people, and is always hopeful about the outcome of situations. As such, for the purposes of this example, the optimist is taken as the agent who always selects the maximum trust value from the range of experiences he has had. Consider figure 4.1. The optimist will take the peak figure in this distribution, hence in formula 4.1, we would substitute that value for $\widehat{T_x(y)}$. There are other considerations to make regarding optimism and the other dispositions discussed here, particularly with regard to alteration of the general trust value for an agent. We present these below.

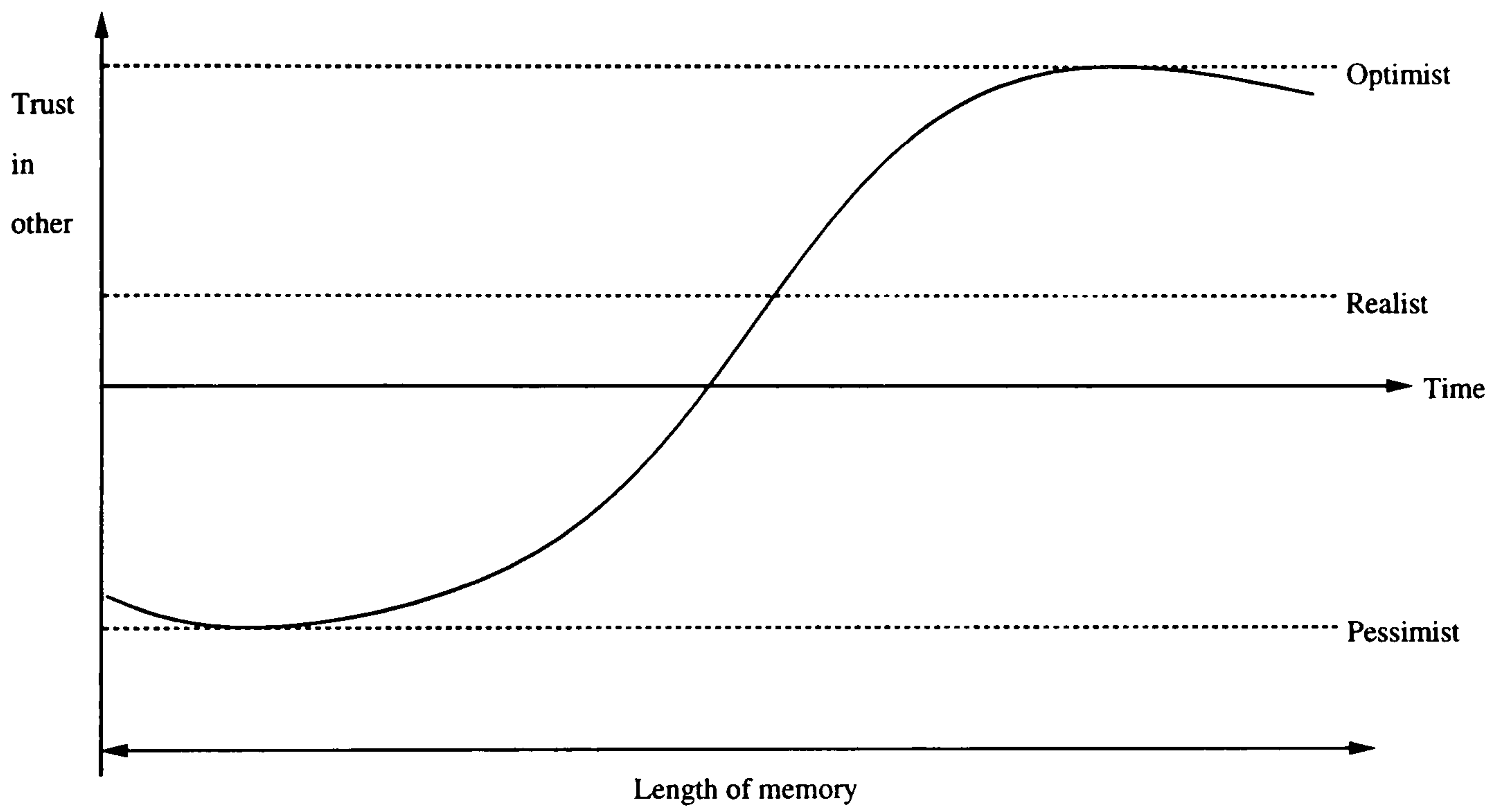

Figure 4.1: Fluctuations in trust for a typical agent, showing the possible estimates used for determining situational trust in other.

4.6.2 Minimum Estimate — Pessimism

In contrast to the optimist, the pessimist is one who sees the worst in people, always doubting the result of situations. It seems clear, then, that the pessimist will take the worst possible value in all of the data he was presented with. In figure 4.1, the pessimist would take the lower value indicated, substituting this in formula 4.1 for $\widehat{T_x(y)}$.

4.6.3 Pragmatism and Realism

It should be clear from the above discussions that, although there are extremes of optimism and pessimism, most of us would consider ourselves to be 'somewhere in-between.' Indeed, it may be the case that, depending on the situation, or even the side we got out of bed that morning, we may choose (consciously or otherwise) to be overly pessimistic or optimistic. This suggests that there exists a spectrum of dispositions within which we exist. This spectrum is shown in figure 4.2. The same spectrum will apply to our agents, but for simplicity, we take two of the points along the spectrum, and discuss them here.

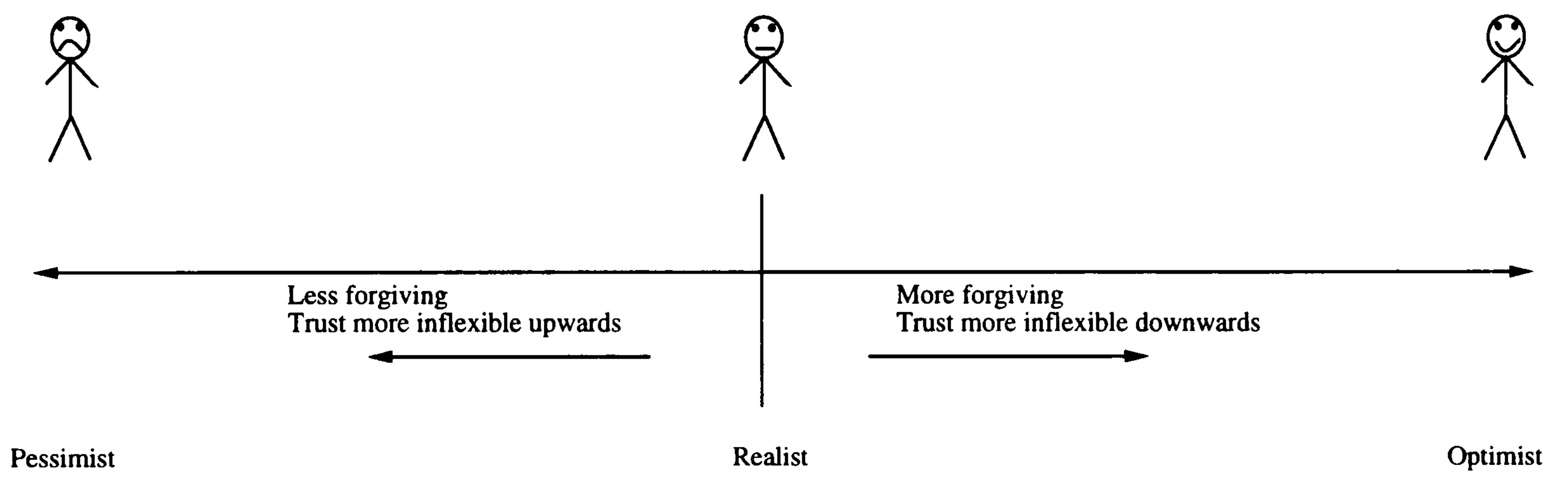

Figure 4.2: Possible spectrum of behavioural dispositions (from Marsh, 1994).

The mean is an example of a point along the spectrum which is relatively easy to obtain values of $T_x(y)$ for. It was first presented in (Thimbleby *et al.*, 1994) and is discussed here due to its practicality as a solution to the problem of estimating the trust value, but one which is perhaps less susceptible to memory lapses and forgetting than the extremes of optimism or pessimism (For a discussion of memory within agents, see later in this chapter). To work out the mean, we use:

$$\widehat{T_x(y)} = \frac{1}{|A|} \sum_{\alpha \in A} T_x(y) \tag{4.2}$$

Here, A is the set of situations similar to the present situation (α) which x has experienced *with* y. This is bounded in two ways:

- x may not have known y for long, thus will not have a large sample size to draw on;
- There will be a bound on memory which will allow x to remember only so far back with respect to situations (see below).

Another consideration here may be that x would want to consider all situations within which he has interacted with y, or more expansively, all situations he has ever experienced, either with y or not. These considerations are in fact too extensive in practicality, and with memory bounds, will only provide an agent with information of limited utility. The more restrictive estimate, using similar situations, is also more sensible and of more use, since the information it gives the agent pertains to situations of a similar nature to the present one, which is in the spirit of situational trust. We remain with the more restrictive implementation here (and in the practical experiments — see chapter 7).

The mean represents what might be termed a 'pragmatic' or 'realistic' point on the spectrum in figure 4.2. It is not, in other words, pathological or extreme behaviourally.

4.6.4 Comments

Final experiments with this set of formulæ are shown in chapter 7, along with discussions of how the various strategies available to the trusting agent affect its overall decisions.

The strategies presented here are not the only ones available. Another option for the trusting agent may be to take the mode of all $T_x(y)$ in the past. Yet another may be to substitute the final situational trust, $T_x(y, \alpha)$, for all situations known with that agent in the formula for $\widehat{T_x(y)}$. Another would be to find the mid point of all $T_x(y)$, and so forth. Much remains to be done to find the ideal measure for the trust estimate. We remain with optimism, pessimism and realism for a number of reasons:

- Optimism and pessimism provide ideal examples of extremes of behaviour. It is of interest to see how much they differ from each other in the final analysis, and how much they differ from realism.

- The realist, or pragmatist, disposition forms a control point from which to study the behaviours of many other dispositions, including those mentioned above.

- The realist disposition is also of merit by itself, since, as the results from experiments show (chapter 7), it holds some surprises of its own, particularly in terms of being affected by extreme points in the agent's memory of situations (see section 7.4.2 in particular).

- The range of possible dispositions is potentially infinite, spread along the spectrum shown in figure 4.2. We have chosen some samples along that spectrum which we believe provide a good insight into the workings of trust and how it is affected by an agent's disposition.

4.7 The Cooperation Threshold

Having determined the situational trust in y, x can consider whether or not to cooperate. In other words, does x trust y enough in this situation to cooperate with her? To determine the answer to this question, we introduce the concept of threshold values for trust. If the situational trust is above a *cooperation threshold*, cooperation will occur, if not, cooperation will not occur.

This view of cooperation may seem somewhat simplistic. Although it is extended further in the following chapter (see specifically section 5.2.3), it presents a straightforward notion of cooperation which is deeper than many (e.g., Tit for Tat (Axelrod, 1984; Godfray, 1992; Novak and Sigmund, 1992), which simply uses a memory span of one action to determine its next reaction) and less deep than others (e.g., many of the strategies used in Axelrod (1984), also planning architectures, such as APE (Spector and Hendler, 1990), and GOFER (Le Pape, 1990); see Ferguson (1992) for a detailed discussion). In addition, it provides an interesting view of the problem which is not so complex as to put substantial overheads onto an implementation (see chapter 7). Also, the notion of trust presented here is not intended to be the sole decision strategy in use in an agent at any time. It is intended to be an augmentation to the agent's current reasoning strategy. A trusting Tit for Tat player, for example, may be more successful in Axelrod's experiments than simple Tit for Tat. For this thesis, trust is considered in isolation for clarity. So:

$$T_x(y, \alpha) > \text{Cooperation\_Threshold}_x(\alpha) \Rightarrow \text{Will\_Cooperate}(x, y, \alpha)$$

4.7.1 Determining the Cooperation Threshold

> ...the *optimal threshold* of the probability of believing we trust someone enough to engage in such [trustworthy] action will not be the same in all circumstances ... we can not only expect the threshold to vary *subjectively*, as a result of individual predispositions (one's inclination to take risks or

> degree of tolerance of potential disappointment); we can also expect it to vary in accordance with *objective* circumstances. For example, it will be higher when the costs of misplacing trust are potentially higher than those of not granting it at all and refraining from action: to walk about in a trench in sight of the enemy (to pick up the example discussed by Axelrod (1984)) requires an extremely high degree of trust that the enemy will observe the implicit truce, and the costs of being wrong may prove much more serious than those of lying low.
>
> (Gambetta, 1990a, page 222)

The cooperation threshold is a subjective measure, tempered by objective beliefs. It is calculated here in a similar fashion to the caluclations for situational trust. In earlier work (Marsh, 1992; Marsh, 1993) we suggested that the cooperation threshold was determined by:

$$\text{Cooperation\_Threshold}_x(\alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y,\alpha)} \times I_x(\alpha) \tag{4.3}$$

This was thought to be too restrictive, as it was felt that it did not take into account the importance of trust in its determination. In addition, the notion of risks over competence is too shallow without taking trust into account (competence in particular is a shallow notion which suffers from its agent-centredness, or subjectivity). We therefore in the later stages of this work introduced an alteration to this formula, which is now:

$$\text{Cooperation\_Threshold}_x(\alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y,\alpha) + \widehat{T_x(y)}} \times I_x(\alpha) \tag{4.4}$$

Where $\widehat{T_x(y)}$ is as discussed above. Here, trust plays a role in the mediation of the cooperation threshold — a very low trust will ensure cooperation is less likely to occur than if trust is high. This is in keeping with psychological findings (Rempel *et al*, 1985; Rempel and Holmes, 1986).

The cooperation threshold formula is as important as that for situational trust (formula 4.1). We therefore present a discussion of the formula in table 4.4, with associated notes. In the event, the incorporation of the value of $T_x(y)$ in the formula poses some problems, and may well be taken out again in further work. For the remainder of this thesis, however, it remains, since the examples chosen are not seriously affected by it, and because its inclusion may provide insights into the working of trust in cooperative situations.

		Perceived_Competence$_x(y, \alpha) + T_x(y)$					
		-1	-0.5	0	$+0.5$	$+1$	$+2$[a]
	0	0[1]	0[1]	∞[2]	0[1]	0[1]	0[1]
Risk [b]	$+0.5$	-0.5[3]	-1[4]	∞[2]	$+1$[5]	$+0.5$[6]	$+0.25$[7]
	$+1$	-1[8]	-2[9]	∞[2]	$+2$[10]	$+1$[11]	$+0.5$[12]

Table 4.4: Examination of the formula for determining the Cooperation Threshold.

Notes on table 4.4.

[a] Since the Competence can only be zero or greater, there is no -2 to mirror the $+2$ here. Actually, although neither is possible, it is sensible to examine the extremes.

[b] Perceived Risk can only be greater than or equal to 0 also. A negative risk is not sensible here. This column measures Perceived_Risk$_x(\alpha)$.

[1] This is sensible behaviour: if no risk is involved, then the Cooperation Threshold can be effectively ignored. Note that this is only one way of addressing the problem: another would be to concentrate more on trust or competence here to arrive at a suitable estimate. This is an avenue for further work.

2 This occurs when Perceived_Competence$_x(y, \alpha) = -T_x(y)$. This is a problem with the representations used here, and the operations upon them. It suggests also that the removal of the trust value from this equation may be a sensible choice in the future. Thus, in further work, we may return to the original formula for the Cooperation Threshold, without consideration of trust.

[3] This behaviour is clearly nonsensical. Low competence and high risk should sensibly give a higher Cooperation Threshold, which is not happening here. Again, this problem arises because of the presence of the $T_x(y)$ value in the formula.

[4] This is again not sensible. See note 3.

[5] (With notes 6 and 7) As risk remains static, an increasing competence decreases the Cooperation Threshold. This is sensible and desirable behaviour in a rational agent.

[6] See note 5.

[7] See note 5.

[8] This behaviour is againnot sensible. See notes 3 and 4.

[9] See notes 3, 4 and 8.

[10] High risk with low competence, and therefore a high Cooperation Threshold. This is sensible, but does lead to artificially high Thresholds, which may be sensibly normalised in some way.

[11] Again, this is sensible behaviour.

[12] This is sensible. Taking notes 10 and 11, the resultant behaviour is as could be expected, sincce as trust and competence increase and risk remains static, the Cooperation Threshold decreases. Taking notes 1 and 7, as competence and trust remain static and risk increaes, so does the threshold.

4.7.2 Too Important, or not Important enough?

Formula 4.4 presents one view about cooperation: this is that the more important a situation is, the more we need to trust someone to enter into cooperation with them in that situation. Hence, as $I_x(\alpha)$ increases in the formula, so does the cooperation threshold.

There is, however, another possible view, which characterises a different way of thinking about things. This is that the more important a situation is, the more we need to get it done, so the lower the threshold of cooperation should be in order to guarantee it getting done.

Both of these considerations have their place; both are equally valid. They may even be adopted by the same agent in different situations. The second of them may signify that the agent concerned has critical time constraints on the completion of the situation. Hence, the importance is associated with temporal considerations. The former consideration may signify an agent who wants something done extremely well, and thus the importance is based on satisfactory completion of the situation without time constraints. There are other such examples, notably that the former consideration signifies an agent who has little confidence in the abilities of others, despite how much he might trust them.

In fact, where importance is seen as a function of time, it is plausible to consider it to relate closely to the concept of 'urgency' for an agent: where it is important to get the job done quickly, then the job is urgent.

In earlier work (Marsh, 1992) we took the former consideration to be paramount. and we will tend to remain with this formula for the remainder of this thesis. Thus, the equation for the cooperation threshold will mostly remain as in formula 4.4. However, there may be situations in which the latter consideration becomes of interest. Then the following formula could be used:

$$\text{Cooperation\_Threshold}_x(\alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{(\text{Perceived\_Competence}_x(y,\alpha) + \widehat{T_x(y)}) \times I_x(\alpha)} \tag{4.5}$$

It is problematic and paradoxical that, for identical importance values, this formula will result in higher cooperation thresholds than formula 4.4.

The inclusion of importance, coupled with the mechanism we use for determining situational trust, gives us a powerful decision making tool not limited to simple probabilities, but relying also on agent-centered measures or estimates. The main contribution of such a subjective measure is that is is based on experience of that agent in the past, and is flexible enough to incorporate the changing requirements of the environment in which the agent is embodied. What was not very important last week may well be of vital importance this (for example, revision for an exam tomorrow, or delivery of aid to a crisis area which has just flared up). The inclusion of subjective or agent-centered measures allows powerful, flexible, trust-based decisions to be made.

4.7.3 Risk

Risk is an important component of trust:

> The incorporation of risk into [a] decision can be treated under a general heading that can be described by the single word "trust." Situations involving trust constitute a subclass of those involving risk. They are situations in which the risk one takes depends on the performance of another actor.
>
> Coleman, 1990. Page 91.

That risk plays such a large part in trust is clear. How to estimate the potential risks in a situation is not: "Decisions involving risks illustrate the limits of human rationality..." (Zeckhauser & Viscusi, 1990, Page 559). Zeckhauser and Viscusi discuss the difference between risk, uncertainty, and ignorance: "in the situation of risk, we know the states of the world that may prevail and the precise probabilities of each state. In the case of uncertainty, the precise probabilities are not known. With ignorance, we may not even be able to define what states of the world are possible." (*ibid.*, page 561). The trusting agents we consider may well be operating under conditions of uncertainty or ignorance. As time progresses with similar situations, experience will give them more of a means of determining risk accurately. This suggests that the determination of risk involves differing methods for differing situations. We will consider three states for a given agent considering the risks involved in a particular situation, β:

1. The agent has no knowledge or experience of β.

2. The agent has some incomplete knowledge or experience of β.

3. The agent has considerable experience or knowledge of β.

State 1, where the agent has absolutely no knowledge of the situation, is on the face of it the more problematic of the three. The agent is acting under a state of ignorance, and any decision could be wildly wrong. We suggest a strategy which involves the use of how much the agent trusts itself *within the situation of making such a decision*. For example, if the situation of ignorance is characterised as δ, then the agent, call it z, will take into consideration $T_z(z, \delta)$. Past experience in the decision situation (δ) will allow z to put some form of certainty on the risks it thinks may be involved with β. Of course, they may be erroneous, but then z will trust itself less in making such decisions in future, and may look further at a problem before jumping into it. In other words, it will become more cautious in unknown situations.

State 2, the state of uncertainty, implies that the agent knows the possible world states that may arise from the situation. It may also know some of the probabilities of such states, in which case it is in a fairly good position, being an adaptive, learning agent (Zeckhauser & Viscusi, 1990). If none of the probabilities are known, then the procedure for state 1, involving the agent's trust in itself, will again prove useful, providing a measure of certainty in the probabilities it ascribes to the different states.

In **State 3**, the agent will simply assess the risks involved, perhaps using a form of Bayesian theory (see Zeckhauser and Viscusi (1990) for an example), arriving at some usable metric.

This discussion of risk provided few concrete answers as to 'how to calculate the risk inherent in a situation.' It did, however, provide some pointers which can be useful. In the state of risk with 'complete knowledge,' the answer is clear, and the agent knows how to estimate the risk (i.e., take the mean assessment of the probability of each outcome). The other two states require some initiative on the part of the agent. Any mistakes made will be learnt from, however. It is primarily through experience that the information necessary to assess such risks and to assess one's own capabilities as regards that assessment, is obtained.

4.7.4 Competence

Competence, as risk, involves an agent making a judgement about someone who may or may not be known to the agent. Once again, there are three possible states to consider. Here, however, additional considerations are available which reflect what society can do to help for an individual. They concern the law, and professional status and societies. In order to understand the choices an agent can make regarding competence, it will be useful to discuss membership of professional societies beforehand. Later considerations will involve the legal aspects of contracts, and so forth.

4.7.5 Membership of Professional Societies

Whether or not an agent is known to our truster, it may be that the agent has another way of securing cooperation, of guaranteeing 'good work done.' Society has a legal system which ensures that certain 'bad' things should not be done (Lagenspetz (1992), see also section 5.3 below). This system, however, sometimes does not stretch as far as it could. Take an example of furniture removal; here, we expect our furniture to arrive unscathed at its destination, to be treated well. There isn't much we can do about it if it doesn't unless the removal man is a professional — a member of a professional institution. In this case, we know that the institution has set down guidelines (ethics or laws) which its members must follow. Our furniture *should* be treated well.

This example introduces the concept of professionalism. Professionals become members of professional institutions, and are acknowledged by these institutions, to assure the public that they will do their job properly, an approximation of Barber's *fiduciary responsibility* (Barber, 1983). In computing, the situation is similar (Association for Computing Machinery, 1992). Being a member of such an institution means that people accepting your help or work will expect you to adhere to the ethics of that institution or face the consequences. Your competence will be judged on this score.

Membership of a professional institution begs the question of whether or not trust is needed any more. This, along with what legal restrictions society can impose on individuals is discussed in section 5.3.

We proceed with our discussion of competence.

4.7.6 Three States of Competence

As with risk, there are three states of knowledge regarding the competence of an agent under consideration. They are:

1. The agent is not known, in this or similar situations ($\neg K_x(y)$ holds $\forall \alpha \in \mathcal{S}$).
2. The agent is known, but not in this or similar situations.
3. The agent is known and trusted in this or similar situations.

These states are discussed below:

State 1: The trustee is not known

This is perhaps the most problematic situation, since in the basic situation there is no way of judging the other. In a situation where absolutely nothing is known, a sensible measure to use might be the general trusting disposition of the truster, moderated by how important the situation is. Once again here, we have to consider whether the agent believes the situation is too important to get done badly, or too important not to get done at all (see above). In contrast to the cooperation threshold, we consider that, if the situation becomes more important, whatever we may want in terms of getting it done, we should expect a more competent agent to do it, since, should we want it done quickly, the more competent agent will do it, and should we want it done well, then, again, the more competent agent can do it. There are alternative considerations (e.g., who cares about the competence as long as the job gets done). We do not consider them to be viable alternatives since we make the assumption that our agents will expect competence no matter what. We use the following formula in this instance:

$$\text{Perceived\_Competence}_x = T_x I_x(\alpha)$$

State 2: The trustee is not known in this situation

In this case, the trustee may be a friend who offers to help, or who requests help, in a situation within which both agents have not yet interacted. Thus it is impossible to use past actual competence values *in similar situations* for the competence of the trustee. The truster, however, will know of the actual competence of the trustee in *some* situations (at least one, otherwise we are in state 1). Thus we can take this into account in some way. We suggest that this knowledge is moderated by the amount of trust the truster has in the trustee in general. So, the following is used to determine competence:

$$\text{Perceived\_Competence}_x(y,\alpha) = \frac{1}{|A|}\sum_{\beta\in B}(\text{Experienced\_Competence}_x(y,\beta)^{t'}) \times \widehat{T_x(y)}$$

Here, $\widehat{T_x(y)}$ is as in formula 4.1, and β is the set of *all* situations in which x has had interactions with y. By definition, these will all be dissimilar to situation α.

State 3: The agent is known and trusted

Since competence naturally takes into account the agent to be trusted, the problem of ascribing competence is relatively small when the agent is known, and smaller still when the situation is similar or identical to one in which the two agents have interacted in the past. The considerations are clearly agent-centered. A simple means of estimating competence is looking at similar situations in the past (here we come back to memory span, for which, see below). One solution is to take all of the resultant competence values from these situations, and to gain an estimate of the competence from these — after the situation has completed, we *know* the value of the competence of the agent, or at least, the competence of that agent at that time in that situation. Taking a statistical measure of all of these competence values helps eliminate the problems associated with, for example, our only knowing similar, not identical situations. It helps reduce the amount of error. This brings us back to which measure to choose (e.g., optimism, pessimism or realism). We will use realism (the mean) here. So:

$$\text{Perceived\_Competence}_x(y,\alpha) = \frac{1}{|A|}\sum_{\alpha\in A}(\text{Experienced\_Competence}_x(y,\alpha)^{t'})$$

Where t' represents the fact that the competence value is taken *after* the situation has concluded. It is thus an accurate representation of what was experienced in that situation.

Professional Societies

In all situations, there may be another way of judging competence, such as membership of a professional institution (see section 4.7.5) or other qualifications. There may also be other considerations, such as money withheld for a poor job done (section 5.2.4), or a legal binding (section 5.3). In all of these cases, the competence may be relatively simple to determine. In addition, we need not consider the competence of the individual at all, rather what we know of the organisation he is a member of.

4.8 Memory

Since time is an important aspect of the real world (Coleman, 1990), it follows that we want our agents to be able to consider temporally as well as socially, and, perhaps, spatially. One of the benefits of trust is that in one value, a large amount of past experience can be represented. That is, since the trust in an agent will rise and fall according to the past experiences the truster has had with that agent, the final value is a good approximation of how well the trustee *should* be trusted.

There are, however, other considerations. When determining situational trust, for example, we would wish our agents to make as informed a decision as possible. Clearly, the more information regarding trust values in the past that the agent has, the better informed it will be; thus we incorporate the concept of memory into a trusting agent to allow deeper consideration of past events, enabling greater predictional power. The aspects we would like our agent to remember as far as trust is concerned are relatively simple: the trust values for past events, for example. From an implementation point of view, these can be stored in very little space within the agent. Space, however, is not infinite. Memory is therefore a finite concept for trusting agents. In reality, a bound will be placed on the amount of situations the agent can 'remember.' A simple boundary value is time — we say that the agent can only remember so far back in time, and can set that value to whatever we wish. This allows us in an implementation to study the effects of short memories or long memories in agents of particular dispositions (see above). Another bound is the number of interactions with particular agents; for example, an agent can remember the last 10 interactions with all agents. For more realism, this could also be bounded by time, such that if the last interaction x had with y was, say, 10 years ago, x would have trouble recalling it unless it was meaningful.

4.8.1 Memory Span

An agent has a memory span, which is the number of things it can remember, bounded by time (or some other realistic metric). The memory span of an agent is signified by Θ, with $0 \leq \Theta \ll \infty$. As an example, consider the formula for taking the mean of the available trust values in the determination of situational trust (see formulæ 4.2 and 4.1). The formula states:

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

For agent x, A is the set of all situations similar to the present one within which he has interacted with y. The bounds of this set are determined by Γ for x (call this Γ_x). So, for example, x will not remember a situation at time δ, if $\delta < \Gamma_x$. If the number of interactions is the bound, the result is similar, with x not remembering more than 10 interactions back.

4.8.2 Problems with Memory

This implementation is a simple one. It presents us with some problems, however. Most notable is the following. Consider figure 4.3 for x estimating values for y. The peak value of trust at $t1 - m$ will be taken into account for all of the methods we describe above for obtaining the estimate of trust. Thus, an optimist would choose a very high value here. The pessimist would remain low, and the mean and mode would take their respective, less sensitive, values. If x has a memory span equal to m, however, a major problem occurs on the next time round, particularly for the optimist. The peak is not taken into account — it is 'forgotten' by x. This clearly has ramifications for the estimate of trust; it is reduced by a large amount for optimists, at least. The mode would not be affected by such peaks, however, and neither would a pessimist be affected, although the same effect would be seen with a deep trough in the graph. It is worth noting that such peaks and troughs should generally not occur — the amount of trust will be a gently fluctuating value which changes only slightly over a short period of time. Rarely do we completely lose trust in someone we trusted to a large extent yesterday.

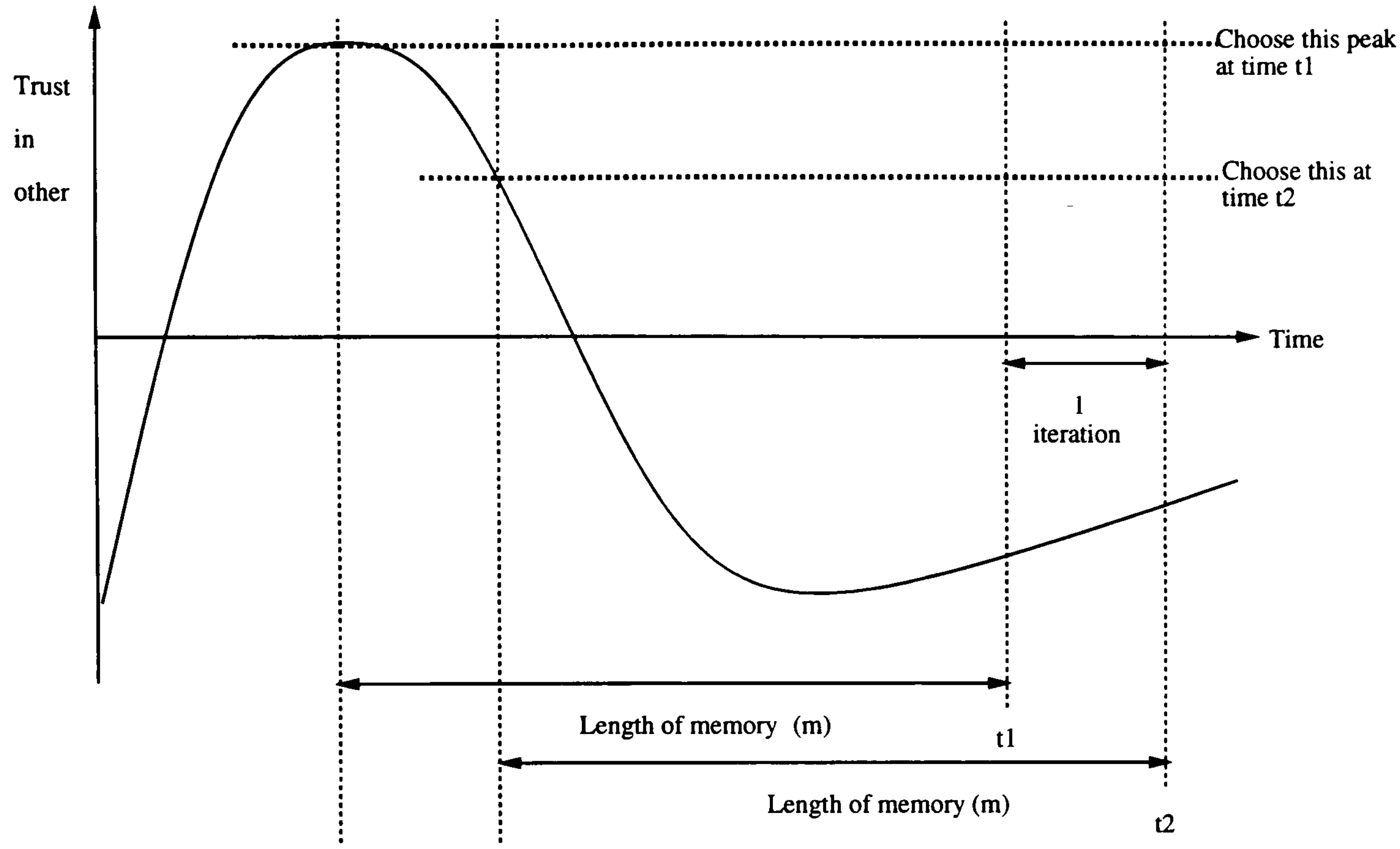

Figure 4.3: The effects of memory on decisions involving trust.

4.9 Reciprocation

4.9.1 Discussion

For the purpose of defining trust, our examples make the assumption that the only considerations agents make involve trust. They should also, however, have some knowledge of certain real-world concepts. Perhaps the most important of these concepts is that of reciprocation — if one does another a favour, one expects that favour to be returned, either now or at some time in the future. Reciprocation is a common phenomenon — bats are known to reciprocate (Harcourt, 1991), as are chimpanzees and cetaceans, amongst others (Trivers, 1985). Indeed, delayed reciprocation in animals can be seen as a precursor to, if not synonymous with, trust (Harcourt, 1991). The notion allows an extension to reasoning about trust, since some expectation of future actions can be based upon expected present behaviour as a result of past behaviour. For example, if x had helped y with homework last week, she might reasonably expect y to help her move this table today. Reciprocity is also likely to benefit the reciprocator in the long run too: "Reciprocity is [an] important source of non-random social interaction... If individuals commonly employ some such rule [of reciprocation], cooperative interactions among reciprocators will persist, while interactions of reciprocators and non-cooperators will quickly cease. Thus reciprocators will be more likely than non-cooperators to receive the benefits of the cooperative acts of others." (Boyd & Richerson, 1991, pages 28–29). Considerations of reciprocation can easily be added to the formalism in order to allow a deepening of the consideration of one agent or situation by another. We provide an example of an extension to the formalism to that effect in the following sections.

4.9.2 Reciprocation — Extending the Formalism

In considerations of reciprocation, aspects of memory, and thus time, need to be addressed. There are several different states of the world with respect to memory that require consideration before we can arrive at a satisfactory representation for reciprocation. We will consider the case of two agents: $x, y \in \mathcal{A}$, where x is considering whether or not to cooperate with y. Assuming that $K_x(y)$ is true, thus there is a calculable value for $T_x(y, \alpha)$, where α is the situation x finds herself in. Also, there are reasonable straightforward means for x to calculate Risk and Competence for y in α. Assuming t_n is the present time, and $t_n - \delta$ is some time in the past (not necessarily identical for each of the following states) we consider the possible states of the memory of x and y:

1. At some time $t_n - \delta$, y has helped x (and it was not a reciprocation for help x had given y at some earlier time), and x has not yet reciprocated.

2. At some time, $t_n - \delta$, x has helped y (likewise, it was not a reciprocal act), and y has not yet reciprocated.

3. At some time, $t_n - \delta$, y helped x, and at a later time, x reciprocated.

4. At some time, $t_n - \delta$, x helped y, and at a later time, y reciprocated.

These states are summarised in the first column of table 4.5, below, as $y \neg x$, $x \neg y$, yx, and xy. There are more things to consider, however; these are summarised in the same table. They are that x and y have finite memories, thus either or both of them may have forgotten the actions carried out at $t_n - \delta$. The final revision of trust for both agents is different depending on what happens (does x decide to cooperate or not?) and what was expected (did y expect cooperation, or had she forgotten that reciprocation was to be expected?) are different for each consideration. Memory in this example is not of the last two interactions for each agent, rather of the last two interactions they had together. For example, during the course of a day, agents may interact with many other agents, and meet up with the same agent they started with at the end of the day. As far as these two are concerned, it is the interactions they last experienced *with each other* that are to be considered. Thus, when we say xy, for example, we mean that the last time these met, y 'helped' x, and that the time they met before that, it was x that helped. Clearly, then, memory is an important aspect of this. The four cases given above do not cover the whole aspect of the problem; it may well be that in xy, y can only remember that she helped x at some time, and thus perceives that this situation is actually $y\neg x$. In which case, looking at the interaction from y's point of view, cooperation is expected, rather than 'hoped for.' The table handles these situations, since we can look up from y's point of view what the situation is. It can only be one of the four, xy, yx, $x\neg y$, or $y\neg x$ in any case.

State	x remembers?	y remembers?	x favours cooperation?	y trust s when	
				x cooperates	x doesn't
$y\neg x$	TRUE	TRUE	YES	$\leftrightarrow$	$\Downarrow$
$y\neg x$	TRUE	FALSE	YES	$\uparrow$	$\downarrow$
$y\neg x$	FALSE	TRUE	SEE TEXT	$\leftrightarrow$	$\Downarrow$
$y\neg x$	FALSE	FALSE	SEE TEXT	$\uparrow$	$\downarrow$
$x\neg y$	TRUE	TRUE	NO[1]	$\uparrow$[2]	$\leftrightarrow$ or $\downarrow$[3]
$x\neg y$	TRUE	FALSE	NO[1]	$\uparrow$	$\downarrow$
$x\neg y$	FALSE	TRUE	SEE TEXT	$\uparrow$[2]	$\leftrightarrow$ or $\downarrow$[3]
$x\neg y$	FALSE	FALSE	SEE TEXT	$\uparrow$	$\downarrow$
yx	TRUE	TRUE	SEE TEXT	$\leftrightarrow$	$\Downarrow$
yx	TRUE	FALSE	SEE TEXT	$\uparrow$	$\downarrow$
yx	FALSE	TRUE	SEE TEXT	$\leftrightarrow$	$\Downarrow$
yx	FALSE	FALSE	SEE TEXT	$\uparrow$	$\downarrow$
xy	TRUE	TRUE	SEE TEXT	$\leftrightarrow$	$\Downarrow$
xy	TRUE	FALSE	SEE TEXT	$\uparrow$	$\downarrow$
xy	FALSE	TRUE	SEE TEXT	$\leftrightarrow$	$\Downarrow$
xy	FALSE	FALSE	SEE TEXT	$\uparrow$	$\downarrow$

Symbol	Interpretation
$\leftrightarrow$	Trust remains the same value (or changes little).
$\uparrow$	Trust increases by some amount, δ.
$\downarrow$	Trust decreases by some amount, δ.
$\Downarrow$	Trust decreases by some amount, γ, where $\gamma > \delta$.

Table 4.5: Possible memory states and outcomes for reciprocation

Discussing the Table

There are several aspects of the table that need to be addressed. Firstly, note 1 suggests that x does not favour cooperation in the $x \neg y$ state. This requires clarification, since it is not always the case. In the context of a single set of interactions, x is less likely to cooperate with y if he perceives y is already in debt to him. There is another consideration, however, in that x and y may be in an ongoing trusting relationship, in which case, reciprocation will be a matter of course, and x will not mind helping again, since some at time in the future, he will expect to benefit from this. Hence, at note 1, we could substitute a YES if the relationship is a continuing one.

Note 2 suggests that y will trust x more should he cooperate in $x \neg y$. This is because y can perceive that x is helping even though she is already in debt to him. Once again, in the context of an interaction, this may not occur, especially if the situation is such that the two are very close, with a high trust value between them anyway.

Note 3 is explained in a similar way. If the two are not close friends in a continuing relationship, y should not expect help from x if she is already in his debt (what reason could x have for helping?). Thus if x does not cooperate, this is no less than could be expected, and her trust in him will stay static — after all, she may have done the same (Lagenspetz, 1992). In a continuing relationship the sudden refusal to help by a partner results in a drastic revision downwards of the trust one has in the other, however. Thus in the same situation, if y cannot remember the situation, and x does not cooperate, we revise trust down less, since the two are not in such a relationship. This represents one aspect of continuing trusting relationships — a sudden defection does a lot of damage (Rempel and Holmes, 1986; Boon and Homes, 1991).

The other parts of the table (at SEE TEXT) can be explained in a similar fashion. If x and y are in a continuing trusting relationship, x would favour cooperation *anyway*, since his trust in y is likely to be high whether he can remember their last interaction or not. In the xy and yx conditions where x can remember, he would probably favour cooperation more if the continuing relationship condition held, not so much if it didn't.

Modifying the Formalism

In addition to the considerations addressed by the table, there are other questions that may be asked, notably whether x is an altruist (see below), and that, whether or not x favours cooperation, he may or may not cooperate. Why is this? The risk involved may be far too high, or in the situation under consideration, x may perceive y to be absolutely incompetent (cf. my brother flying a plane). Thus, when we talk of x favouring cooperation, we mean just that there is more of a chance that he will cooperate. Thus we reduce the cooperation threshold by some value (say, 10%). If it is initially high, there is still a fair chance that the situational trust will not be enough to justify cooperation. This is a consideration to be discussed or addressed in individual situations. It is, however, worth noting that a lowered cooperation threshold and a high trust (as with a continuing relationship) will almost certainly guarantee cooperation.

Thus, the cooperation threshold will be determined by:

$$\text{Cooperation\_Threshold}_x(y,\alpha) = \left(\frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y,\alpha)+\widehat{T_x(y)}} \times I_x(\alpha)\right) \times \text{Reduction\_Percentage}_x(y,\alpha) \quad (4.6)$$

Where the Reduction_Percentage is calculated by x according to the past history of the relationship, as partly described by table 4.5.

Modifying Trust

Given the extra reasoning power from the above formula, the examples below will allow agents to take reciprocation into account both before and after the event. That is to say, for example, if y refuses to help x, then it would be reasonable to expect x to reduce her trust in y by a larger amount than if y had not helped with the homework. This allows us to modify the formalism for increment/decrement of the trust value:
If:

$$\text{Helped}(x,y,\alpha)^{t-\delta} \wedge \text{Defected}(y,\beta)^{t} \quad (4.7)$$

Then:

$$T_x(y)^{t+1} \ll T_x(y)^{t}$$

Informally, if x helped y in the past, and y responded at this time by defecting, the trust x has in y will reduce by a large amount. The converse is if:

$$\text{Helped}(x,y,\alpha)^{t-\delta} \wedge \text{Cooperated}(y,\beta)^{t} \quad (4.8)$$

Then:

$$T_x(y)^{t+1} \geq T_x(y)^{t}$$

Informally, if x helped y in the past, and y reciprocated at this time with cooperation, then the amount of trust x has in y will remain the same or increase only by a small amount.

In other words, the amount of trust x has in y substantially decreases following y not reciprocating (Boon and Holmes, 1991). However, y's reciprocation merely confirms to x that she (x) was correct in helping y in the first place (Lagenspetz, 1992). This being the case, x had every right to *expect* y to help. So, although y's reciprocation may lead x to trust her *judgement* of people more, she may revise her trust in y only slightly, if at all (Lagenspetz, 1992).

The amount of alteration of the trust value at any particular time is a subject of great importance, since it is this which underlies the adaptive nature of trust. However, this work is concerned with the introduction of the formalism for trust itself, rather than its alteration. In the practical work that follows, we have used a simple metric, involving taking or adding percentages to trust (see chapter 7). This is unsatisfactory in the long run, and the effects of different strategies for the alteration of trust are an important aspect of further work.

4.10 Summary

> Trust has never been a topic of mainstream sociology. Neither classical authors nor modern sociologists use the term in a theoretical context. For this reason the elaboration of theoretical frameworks, *one of the main sources of conceptual clarification*, has been relatively neglected.
>
> Luhmann, 1990, page 94. My emphasis.

In discussions of trust (Barber, 1983; Gambetta, 1990b; Gambetta, 1990a; Yamamoto, 1990) there remains an ambiguity associated with the concept. As Luhman notes, this leads to problems when we want to say what trust really *is*! A theoretical framework, based on trust, would give sociologists and social psychologists a tool for precise discussion of the concept. Distributed Artificial Intelligence is a field in which sweeping generalisations can be made regarding social phenomena such as trust. Indeed, trust specifically is glossed over with cursory statements (such as those found in Rosenschein (1985) and von Martial (1992)). It seems, then, that trust is to be considered present whilst not being *really* considered at all.

The work presented in this chapter goes some way to correcting the situation in sociology, social psychology, and DAI. A formalism has been presented which takes into account many of the aspects of trust. It allows precise reasoning about the concept while being relatively simple, and thus easy to implement with very small overheads. One of the principal concerns here was that, with so many agent architectures (Hanks *et al.*, Winter, 1993) and agent description languages[7] the incorporation of trust should be a painless affair. With small overheads in terms of space and implementation, this formalism may satisfy that condition. The evaluation of this is a topic for further work.

The formalism is extensible. There are considerations which have not been made as regards trust here. The following chapter extends the formalism further to take some of these aspects into account, such as choice of who to cooperate with, the law, and how legal aspects in society can help agents in making trusting decisions.

[7] For example, the Michigan Intelligent Coordination Experiment (MICE) language, or MAGSY (Fischer, 1993). See also Hanks *et al.* (1993) for a review of some of the testbeds that are in use today.

Chapter 5

Using the Formalism

5.1 Discussion

Chapter 4 discussed the concept that trust can be formalised, and provided a formalism and associated formulæ which can be built upon to allow reasoning with and about trust. The use of a single final value for trust allows qualitative and quantitative comparisons of different agents' trust values, and with an allowance for trust being a subjective measure, it allows us to use trust as a reasoning tool in embodied agents (see chapter 7 for some implementations of these) and as an analysis tool.

In this chapter, the concept of using trust as an analysis tool and as a descriptive tool for interactions is presented. Several examples are discussed, with particular attention focused on applications for DAI. The examples are mostly from potentially cooperative situations, since, as was discussed in chapter 1, much of the work in DAI is concerned with circumstances involving cooperation amongst agents (see also Marsh (1992, 1993), Rosenschein (1985), Urzelai and Garijo (1992), and for a different approach to this idea Kuwabara and Ishida (1992)). The chapter focuses on furniture moving as an example, and is developed from Marsh (1992), which in turn used the example given in Connah and Wavish (1990). The reasons for the use of furniture moving as an example are given in section 5.2.1.

5.1.1 Formulæ Used

The formulæ presented in the previous chapter will be used in the examples below. The notation is presented in the previous chapter in tables 4.1 and 4.2. For clarification, the basic formulæ used here are as follows:

Situational Trust

This is determined by:[1]

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

Where we are using the mean as the estimate for trust, so:

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

And A is the set of all situations x knows (can remember).

Cooperation Threshold

We determine this as follows:

$$\text{Cooperation\_Threshold}_x(y, \alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y, \alpha) + \widehat{T_x(y)}} \times I_x(\alpha)$$

We thus assume that our agents take the view that, the more important a situation is, the more they need to trust in order to cooperate — situations are too important to get done badly.

The Perceived_Risk is determined by necessarily agent-centered, or subjective, means. See the previous chapter for a discussion of these. Thus, when we come in our formulæ below to a consideration of risk, values will be discussed and presented as they become necessary.

Perceived_Competence is also agent-centered. The three states given in the previous chapter provide some solid answer to the question of how to estimate this. Once again, in the examples below, we will present the means of determining competence 'on the fly.'

As was discussed in the previous chapter, reciprocation is an important aspect of trust. We thus include considerations of reciprocation as we proceed in the following examples.

[1] The fomulæ presented here subject trust to certain limitations. See the discussions in the previous chapter, also section 9.6.

5.2 The Furniture Removers

5.2.1 Why Furniture Removal

The domain of furniture removal is a constrained example of what could, in future, be expected from embodied agents — the sort of work which is heavy, messy, and potentially physically risky (Ferguson, 1992). For more physical risk, there is always work in nuclear installations, or in space. (As an aside, enlightening presentations of this concept, albeit in idealised forms, are to be found in Isaac Asimov's work involving the robots. See in particular "I Robot"). In addition the domain provides an example of an inherently cooperative environment. There will be situations where agents (human or otherwise) cannot manage to move a piece of furniture alone — it is too heavy, or has to be lifted through an awkward space, for example. In such situations, agents will have *no choice* but to cooperate with one another (Argyle, 1991). It is thus realistic to study such a domain in the knowledge that the consideration of cooperation is not a futile one, since it allows us to consider trust unencumbered by other external considerations.

Such inherent necessary cooperation could be seen as forced, or impelled cooperation, particularly if there was only one other agent with whom to cooperate. Hence this section develops the furniture moving domain from two agents, each with furniture to move which is too heavy for one agent alone (section 5.2.2), to a situation involving three agents, with two having the task of moving furniture too heavy for them (section 5.2.4). This example gives each of the agents who *need* to cooperate a choice about *who* to cooperate with, and provides us with a view of how such considerations might work out.

5.2.2 Two Agents, Two Pieces of Furniture

In the first example, we have two agents in a room, each with the task of carrying out one piece of furniture (for simplicity, each must take his piece of furniture to the door). The initial configuration of the example is shown in figure 5.1. The agents are labeled y and z, and the situation is labeled β, so from z's point of view, it is in situation β_z, and from y's, β_y. Note that two agents in the same situation may not call the situation the same thing. So z may have nominated this as situation γ. The present choice of labels is for simplicity and ease of understanding, and to reinforce

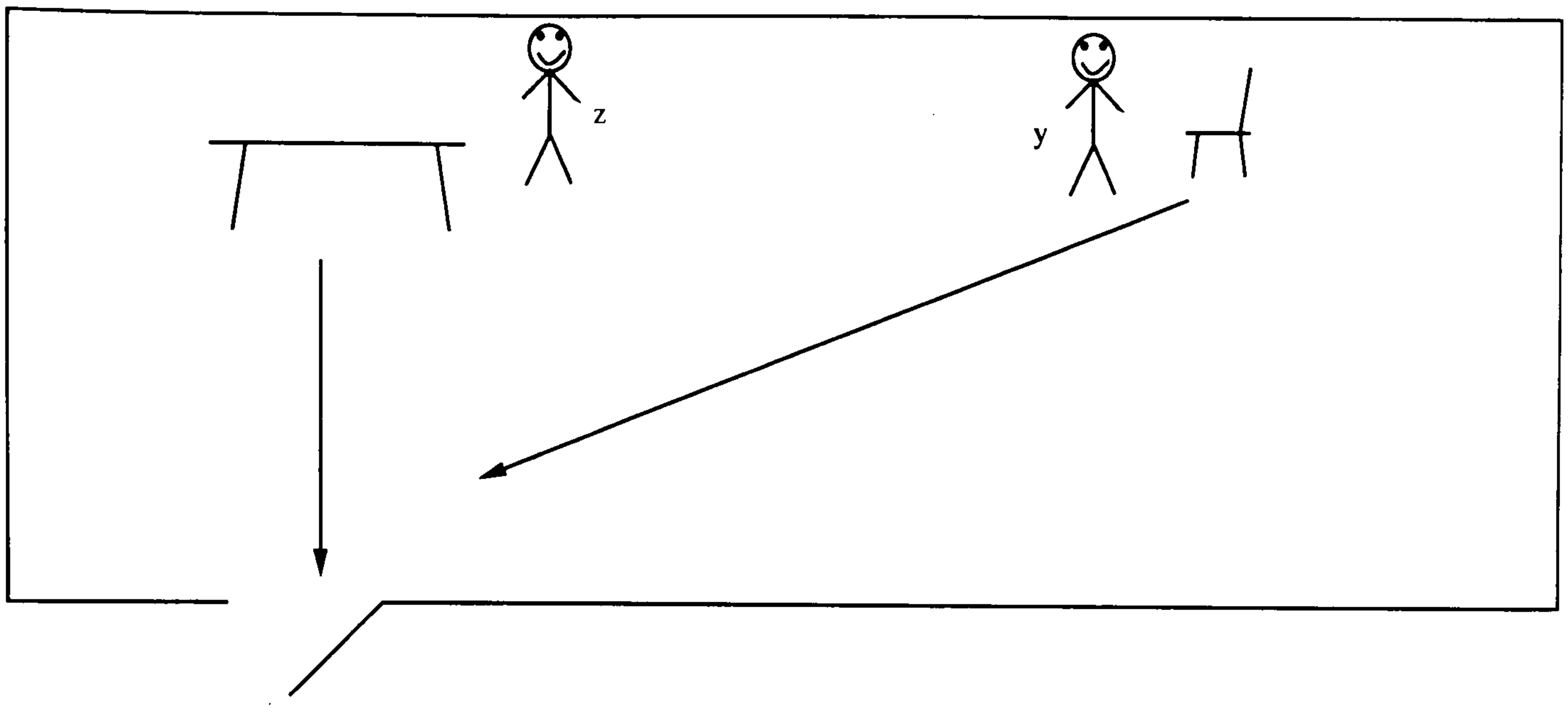

Figure 5.1: Starting situation for two agents with two pieces of furniture to move.

the idea that the two agents are in the same place and time. Further assumptions made are that the agents know each other ($K_y(z)$ is true, as is $K_z(y)$). Moreover, the two have been in several interactions together (hence, estimating competence is not a problem, since the information needed to determine it is available).

The agents in this situation have a problem: each cannot move his piece of furniture alone; indeed each piece of furniture needs two agents to move it, thus each agent must consider cooperation with the other in order to get the job done. For the present situation, consider that they have a choice, they may cooperate and get the job done, or they may choose not to, and nothing will get done. In the second case, there are no sanctions, their orders are not compelling. Deadlock is, then, possible. We consider the situation from y's point of view.

In order to move his piece of furniture, y has to ask for help from z (it may have already been offered). Assuming that, y considers the amount of trust he has in z. Since the two know each other, and have interacted, y can estimate trust fairly easily. The next few sections will present y's considerations. For the purpose of this example, values for the subjective estimates, such as importance, will be presented 'as is.' Since this example is simply to illustrate the working of the formalism, this is not a serious problem. Since trust evolves over a series of interactions, values given without seeing this evolution will seem artificial. The values below are chosen to reflect certain aspects of cooperative situations where trust is involved. In addition, comments about the values below serve to illustrate the example, and are not considered to be part of an agent's 'musings.'

$T_y(z) = 0.6$, hence y trusts z relatively highly in general, but considers the situation to be of fair importance: $I_y(\beta) = 0.5$. The utility y stands to gain from moving the furniture is, he believes, very high, since, for example, the furniture may be his, and he wants it moved from this room. He estimates utility as $U_y(\beta) = 0.8$. From these values, and the values of $T_y(z)$ for as far back as he can remember (these work out as 0.63, thus in the past, he has trusted z more than at present), he can estimate the situational trust he has in z:

$$T_y(z,\beta) = U_x(\alpha) \times I_x(\alpha) \times \widehat{T_x(y)} = 0.80 \times 0.50 \times 0.63 = 0.252$$

This value is on the face of it quite low. This is to be expected, since as was discussed above, the application of the formulæ results in heavily moderated final values. In order to cooperate, or accept z's help, the cooperation threshold must be lower than the situational trust value. This may well be the case. For example, y may consider z to be very competent in this situation. For the purposes of this example, we select values that represent this case. Thus we can show that using trust as a consideration allows cooperation to take place with some certainty as to the outcome. The cooperation threshold for y in β is determined as shown in formula 4.4 in the previous chapter. It uses values for importance, perceived risk, and perceived competence of the other agent in a particular situation. As far as risk is concerned, y can estimate this fairly easily: there is a risk of his piece of furniture being damaged by z's clumsy handling, also of the job not being done in the first place (this is a risk, although it is considered in various other measures, as in the importance). He comes out with a fairly low risk (z is a friend, and he will *try* to be careful — any more than that is considered in the competence value), so $\text{Perceived\_Risk}_y(\beta) = 0.4$. The competence of z is not in question, at the end of the day. Considering all of the furniture moving z has helped him with in the past, y estimates that: $\text{Perceived\_Competence}_y(z,\beta) = 0.7$. Given these values, we can determine the cooperation threshold, which in this case will be:

$$\text{Cooperation\_Threshold}_y(z,\beta) = \frac{0.4}{0.7 + 0.63} \times 0.5 = 0.1504$$

This is indeed quite low. In addition, it is below the situational trust of 0.252. Cooperation between y and z will be possible from y's point of view (although z may have different ideas. We do not consider z or his perceptions in this example). There is one further consideration to make, and that is that the two are in a continuing relationship which involves reciprocation. Whilst this will make no difference to the outcome of the situation here, it is informative to include a consideration of it. Since reciprocation is expected and given, the cooperation threshold is lowered by a suitable percentage, say, 50%. So the final cooperation threshold is:

$$\text{Cooperation\_Threshold}_y(z,\beta) = 0.1504 \times 0.50 = 0.0752$$

Considering the situational trust, cooperation is almost certain to ensue, providing z agrees to it (this is likely, on consideration, if the two are 'friends').

5.2.3 The *Safety Net*

When two agents find themselves in a situation where both need to get something done, such as the two agents in the previous furniture moving situation, there is the possibility of deadlock if neither trusts the other enough to risk cooperation. We can make allowances for this with trust by incorporating some form of *safety net* into the system. Having a specific measure, or value, for trust allows us see exactly how much extra trust is needed in order to get the job done. The previous chapter suggested that the law and legal aspects may help trusters to accept cooperative interactions with trustees in whom they had little trust. (This is further discussed in section 5.3). Here we extend the situation above to allow for a case where y does not trust z enough to cooperate with her, but has little or no choice. In other words, both agents have strict orders to move the furniture, and reprimands will be very severe if the order is not complied with. These facts will be incorporated into the utility, risk and importance of the situations for the agents concerned, so that cooperation may be more likely in any case. This may not, however, be enough to encourage cooperation.

For this example, we consider once more y's point of view. He knows z, and trusts her little: $T_y(z) = 0.2$ (note that he still *trusts* z, and doesn't *distrust* her). The situation is important; the importance is judged by y to be $I_y(\beta) = 0.85$ (more important than the previous example). The utility to be gained from the situation is unchanged at 0.8. The mean value of trust values so far is determined by y to be $\widehat{T}_y(z) = 0.25$. So the situational trust is as follows:

$$T_y(z,\beta) = U_y(\beta) \times I_y(\beta) \times \widehat{T_y(z)} = 0.8 \times 0.85 \times 0.25 = 0.17$$

This is again a low value, despite the high importance and utility values. The cooperation threshold is likely to be higher, however, since nothing else has changed much.

Perceived competence is unchanged at 0.8, as is perceived risk at 0.3. This may seem strange on the face of it, but the competence of an agent need not be linked to how much he is trusted. The cooperation threshold, then, is:

$$\text{Cooperation\_Threshold}_y(z, \beta) = \frac{0.3}{0.8 + 0.25} \times 0.85 = 0.243$$

Paradoxically, this is higher than the previous example. The situation, however, is more important. There are no considerations for reciprocation — neither is in debt to the other, and they are not in an ongoing relationship. Thus the cooperation threshold stays as it is. So y will decide not to cooperate with z here. There is a problem, then, since it must be done! In this situation, y can set up some form of continual assessment or 'safety net' to protect himself from untrustworthy behaviour on the part of z.

The safety net can take several forms, which can depend on the size of the gap between the cooperation threshold and the situational trust, and may depend on other factors, such as other knowledge the truster (y) has about the trustee (z). Safety nets can come in the form of withheld cooperation (Marsh, 1993), such as y refusing to help z until z helps him out, or conditional cooperation (Danielson, 1992a). These can quickly lead to situations more problematical than the one under consideration, however, with both agents refusing to help until the other does! There are other solutions, though, such as moving each piece of furniture a bit at a time towards the door (somewhat childish, but a solution, nonetheless). Since the taking of a risk is a prerequisite to establishing a trust relationship (Deutsch, 1962; Swinth, 1967; Rempel *et al.*, 1985), it may be sensible for y to take a risk and try to establish a trusting relationship (y's perception of trust may be wildly inaccurate, in which case he would be pleasantly surprised). The other agent will have to perceive that a risk has been taken, however, so this may not be feasible. The safety net will go some way towards allowing cooperation, but determining which form it takes is a problem in itself. Table 5.1 suggests some possible safety nets and the problems and benefits of each.

The examples given in the table are not exhaustive. They do provide an agent with some idea of what can be done when cooperation is a necessary aspect of a job. In the final analysis, there may also be recourse to the law for assistance (see section 5.3, below). The idea of social sanctions may be expressed by legal action, but what is meant by the entry in the table is that, in a society, it may only take one defection or untrustworthy action by one member of that society against another for the first

	Possible Safety Net	Problems	Benefits
1.	Help after help given	Deadlock	Easy to set up
2.	Do tasks bit by bit (alternate)	Takes a long time to do task	Almost as easy as 1.
3.	Do tasks together (simultaneously)	May be impossible	Straightforward and intuitive
4.	y monitors progress continuously	Could be futile (esp. in example)	Gets some of the jobs done
5.	Negotiation	As above Deceit	Could be the only solution
6.	Use social threats (e.g., sanctions)	May not help here Takes place over time	Works well in close-knit societies

Table 5.1: Possible safety nets with benefits and drawbacks of each.

to be ostracised (for another perspective on this, see Lagenspetz (1992)), or put out of business (for example, when banks lose the confidence of their customers, many are brought down in the deluge of customers wanting their money right then). When society decides to take collective (non-violent) action against a member, that member is hard put to do more than just 'exist' in the society. Threats of social sanctions are not hollow.

Other entries in the table merit some discussion. The idea of negotiation is not a futile one: occasionally negotiation produces results. There has been much work on negotiation for problem solving (Sycara, 1988; Galliers, 1989; Chang and Woo, 1991; Zlotkin and Rosenschein, 1992), and some work on reaching consensus plans that maximise social welfare (Ephrati, 1992). All of these could be useful in the negotiation stage to help solve conflicts. Where a safety net cannot be constructed, we are in a classic conflict resolution situation (Hinde and Groebel, 1991b). If the agents recognise this, they can take appropriate actions to remedy the situation. We believe that trust can allow interactions before such a situation arises. The safety net using trust is an example of this.

Table 5.1 suggests that the tasks can be completed simultaneously. In the example of furniture moving, both agents could perhaps carry one piece of furniture in one hand, and the other piece in the other hand. This would get the job done in half the time, but may not be possible — it depends on the domain of the problem. In addition, a problem for one agent may not be akin to a problem for another agent, even in the same situation.

So in the worked example above, where y would not ordinarily cooperate but has to, a safety net must be set up. The safety net could be any of the above suggestions, for example that of negotiation (Galliers, 1989). We conclude that with a safety net, the need for which is suggested by a consideration of trust, cooperation *can* be entered into and 'bad' outcomes (such as z running off after y had helped move z's piece of furniture) accounted for and possibly prevented.

5.2.4 Three Agents, Two Pieces

The above examples concerned two agents interacting in a situation where cooperation was desirable or necessary. In the first, it mattered little whether cooperation was coerced (i.e., necessary) or not, since y would have cooperated anyway. In the second, given a choice, y would ordinarily have chosen not to cooperate. There are examples where y would have cooperated even given the chance not to, with a trust so low. These include situations where, for example, y really needed to get the job done in order to start on something more important (like painting the room). Also, y may have considered that z was the lesser of several evils (better the devil you know...), and thus cooperation with z was 'the best that could be done.' This section looks at an example where another agent is involved (call it x). The benefits of examining such a situation are that:

- It allows consideration of situations containing more than just two agents, yet is constrained enough to show the working of the formalism with clarity.
- It gives agents a choice as to who to cooperate with, if at all.
- It allows us to consider reasons why agents cooperate, other than because they need or have to.
- It provides a more realistic view of real world situations, where there may exist many agents (people).

The situation (call it ϵ) is once again looked at from y's point of view. It is initially as shown in figure 5.2. Agents y and z have to move their piece of furniture to the door, and each piece is too heavy for one agent to lift alone. The third agent, x, is standing in the background, with no task at present to carry out. The question for y is whether to ask for x's help, or z's. For this example, some new values for the determinants of situational trust and the cooperation threshold between y and z will be introduced. This is to reflect that the situation is different to what came before. We assume that y and z are the same two agents as in the *first* example, and that they have interacted since that time.

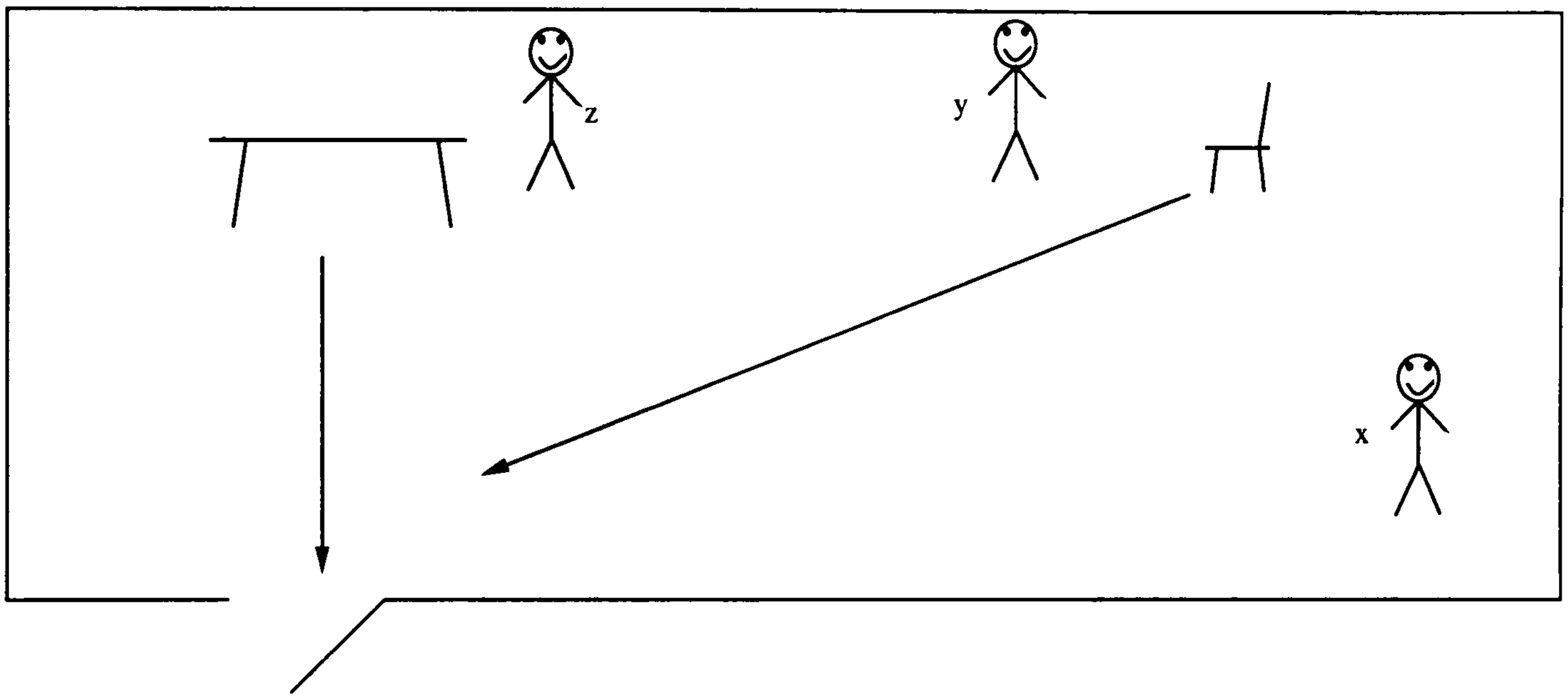

Figure 5.2: Initial starting environment for three agents with two pieces of furniture to move.

y's thoughts on z

These are similar to those given in the discussion in the previous example. In the intervening time, z and y have met once, and both worked together towards a good solution to a problem. The result is that y trusts z more than before, so $T_y(z) = 0.65$ this time around. The situation is as important as in the first example, with $I_y(\epsilon) = 0.50$. The utility to be gained is likewise similar, at $U_y(\epsilon) = 0.80$. Finally, to determine situational trust, $\widehat{T_y(z)} = 0.67$, slightly higher than in the previous example; this reflects the successful outcome of a situation the two have been in since. So:

$$T_y(z, \epsilon) = U_y(\epsilon) \times I_y(\epsilon) \times \widehat{T_y(z)} = 0.80 \times 0.50 \times 0.67 = 0.268$$

This is higher than the previous example, but by a small margin — situational trust takes more into account than the beneficial outcomes of previous situations, and thus is a little less flexible than general trust in that sense.

Working out the cooperation threshold is similarly straightforward for y: he knows the competence of z, and so it remains at $\text{Perceived\_Competence}_y(z, \epsilon) = 0.8$, as before. Risks have changed, though. Since z knows another agent is present, who *may* be able or willing to give help should z fail, y estimates the risk downwards — there is less at stake than before, as there may be a suitable backup should something go wrong (this takes little account of what x wants — see below). Thus, $\text{Perceived\_Risk}_y(\epsilon) = 0.25$. Given the importance of the situation, the cooperation threshold for y and z is:

$$\text{Cooperation\_Threshold}_y(z, \epsilon) = \frac{0.25}{0.8 + 0.67} \times 0.5 = 0.085$$

That this is lower than in the previous example is a reflection of the lower risk involved in this situation. The final threshold is reached after considering reciprocation. As before, the two are in a continuing relationship, so:

$$\text{Cooperation\_Threshold}_y(z, \epsilon) = 0.085 \times 0.50 = 0.0425$$

Were there only two agents in this example, y would now choose to cooperate with z. There is a choice for y here, though. Agent x is in the situation too, so y, realising that things might get better if they were done with x rather than z, considers x also.[2]

y's thoughts on x

For the purposes of this situation, y does not know x ($\neg K_y(x)$), as they have never met before. This leaves y with the problem of assigning suitable values for trust, and so forth, to x. Being a fairly trusting agent, y has a basic trust for $T_y = 0.50$, so assigns this to $T_y(x)$.[3] The situational importance and risk remain the same, as does

[2]This consideration is not strictly necessary or wise, and in a situation where y and z were 'friends,' could be damaging to their relationship — trusted friends do not expect their trust to be questioned (Rempel *et al.*, 1985; Rempel & Holmes, 1986; Boon & Holmes, 1991). However, if y is attempting to maximise utility, there may be the chance that, for example, x can get the job done more quickly. We make the consideration for the sake of the example here.

[3]Here, the agent may substitute 0.00 for $T_x(y)$, since he knows little or nothing of the other agent. This is another reflection of the dispositions of agents. Since y is trusting, he assigns his basic trust value. Experience will lead to the alteration of that value in time if necessary.

the utility to be gained from successful completion. So for situational trust:

$$T_y(x,\epsilon) = U_y(\epsilon) \times I_y(\epsilon) \times \widehat{T_y(x)} = 0.80 \times 0.50 \times 0.50 = 0.20$$

So in this situation, understandably, y trusts x less than he does z. There is an additional decision to be made here, however: x is extremely competent at moving furniture. In fact, he is a professional, and a member of a professional association (and has a certificate to prove it![4]) The competence, then, is estimated by y as $\text{Perceived\_Competence}_y(x,\epsilon) = 0.99$ — very high indeed. The cooperation threshold reflects this:

$$\text{Cooperation\_Threshold}_y(x,\epsilon) = \frac{0.25}{0.99 + 0.50} \times 0.50 = 0.0839$$

This is higher than the final threshold for z (although a little lower than that before reciprocation was taken into account — a reflection of how important reciprocation is). It is, however, lower than the situational trust that y has for x, so y is willing to cooperate with both x and z. This presents another problem for y — who to cooperate with, rather than whether to cooperate or not.

There are several ways of deciding what to do in this situation.

- The most intuitive method is to select the most trusted agent to cooperate with, i.e., if $T_y(z) > T_y(x)$, then y will choose to cooperate with z. This reflects the idea that trust counts for a lot in cooperative decisions.

- It is possible for y to choose the agent who is trusted more *in this situation*, taking the maximum of the situational trusts. Again, this reflects the importance of trust in cooperative decisions, but takes into account how much the situation means to the agent concerned. For different agents, the utility to be gained from the same situation may be different for the truster. For example, a great deal more utility may be achieved if we gain our own utility from that of others (Preston, 1961). Thus the utility we gain from a situation may be different for different agents under consideration in the same situation. This is not reflected in the above example, however.

- Without taking trust into consideration, y could always decide to cooperate with the agent whose cooperation threshold was the lowest. This goes against

[4]See section 4.7.5.

Name	Method	Benefits	Drawbacks
MaxGen	Choose maximum of $T_a(b)$	Easy and quick.	No thoughts re situation.
MaxSit	Choose maximum of $T_a(b, \alpha)$	As above.	No thoughts re competence, etc.
MaxComp	Choose maximum of $\text{Cooperation\_Threshold}_a(b, \alpha)$	Considers risk, competence.	No thoughts re trust.
DiffTrust	Choose maximum of $T_a(b, \alpha) - \text{Coop\_Thresh}_a(b, \alpha)$	Considers whole situation.	Time-consuming.

Table 5.2: Choice Methods: Who to cooperate with?

the spirit of this work, however, and against the view that trust is an essential aspect of cooperation.

- y may choose to cooperate with the agent who has the largest gap between cooperation threshold and situational trust, so if:

  $$T_y(z, \epsilon) - \text{Cooperation\_Threshold}_y(z, \epsilon) > T_y(x, \epsilon) - \text{Cooperation\_Threshold}_y(x, \epsilon)$$

  Then y will cooperate with z, and *vice versa.*

Since for all of these choice methods, y will choose to cooperate with z, this is the natural choice for y. The choice methods are summarised in table 5.2. A final consideration may be to take all of these choice methods, and choose to cooperate with the agent who 'wins' the most.

5.2.5 The Third Man

In the example involving two agents above, both had no choice about cooperation with the other if they were to get their respective jobs done. As was mentioned, this allows us to study an inherently cooperative environment without considering why cooperation takes place, or giving agents freedom of choice over cooperation. The extended example considered three agents, one of which had no furniture to move. As far as we went in this example, we did not consider the third agent's 'feelings' on the matter. The remainder of this section discusses the various options open to this agent.

The third agent has no incentive as such to cooperate with either of the other two — they may ask, and he may help, but he may see no event in the near future that would benefit him, for example by way of reciprocation. It is then possible that he may feel that he is wasting his energy in helping either of the furniture movers. There are explanations for such help if it is given.

- The third agent may be truly altruistic.
- He may simply be reciprocating for some favour one of the movers has done him in the past (which would account for his choice of who to cooperate with quite easily).
- From a possibly more mercenary point of view, he may expect reciprocation of another sort — payment.

If the third man (x) is asked for help by z or y or both, and he is neither an altruist, or in debt and thus partially constrained to reciprocate, then z and y can create a situation which gives x some utility in terms of something x needs. In today's society, this would more often than not mean that x would be paid for his efforts. It is up to x to dictate the cost of his services, dependent on the loss of utility helping z or y would bring. Of course, he may price himself so highly that z and y consider each other once again and decide that, after all, they would be better off helping each other. We consider the possibility that x is either an altruist or demands payment here.

x as an altruist:

If:

$$x \in \{\mathcal{A} : \text{Altruist}(\mathcal{A})\}$$

Then:

$$\text{Will\_Cooperate}(x, a, \alpha)$$

Where:

$$\alpha \in \mathcal{B}_x$$

$\mathcal{B}_x$ is the set of all possible *benevolent* situations for x. A benevolent situation is one in which x can see no possibility of harm to himself, which would result in, for example, death or serious injury. Note that this is not the same definition of altruism that is used in biology: "Biologists define behaviour as altruistic if it favours other individuals at the expense of the altruist itself." (Dawkins, 1989a, page 57). Ours differs from this only in that we restrict the situations to benevolent ones.

Payment to x

If x is not an altruist, and not in debt to either of the other two agents, they may be able to set up a situation where it is worth x's while to cooperate. The easiest way to do this in a world where money is important is to offer payment. This section considers how much x would need to ask for or be offered before he considered cooperation.

Whether or not x trusts y or z, there is a loss to him if he cooperates with either, in terms of utility. How much utility is lost depends on what value x placed on what he could have been doing, or would have been doing, instead of helping the other two. It is thus impossible to determine whether x will help in any particular situation. We can, however, arrive at formulæ which will help x to decide in any situation.

Clearly the utility lost is a significant measure. The importance of the situation is unlikely to be as significant as lost utility since, were it important for x that the situation ended 'well' (i.e., attained an outcome x deemed successful), there should be no question about x cooperating. By definition, then, importance is of little concern in this consideration. Trust in the agent to cooperate with is of importance since x may at some time in the future need help — 'closing doors' unnecessarily is a little rash, and if the other agent is trusted, it may be worth establishing a cooperative relationship, in which case the lost utility is lessened by future expectations (cf. Axelrod's 'Shadow of the Future' in a novel fashion (Axelrod, 1984)). Determination of lost utility is the problem. The following formula is presented as a means of obtaining this:

$$\text{Lost\_Utility}(x, \beta) = \widehat{U_x(\Lambda)} \times T_x(y, \alpha) \tag{5.1}$$

Where $\widehat{U_x(\Lambda)}$ is the mean of all the utilities that x perceives he could have gained from doing other things instead of helping, and $\beta \notin \Lambda$. Clearly this depends on the situation x is in. For example, in the furniture moving considerations, if all x was going to be doing was standing around, then the lost utility may be fairly low. If, on the other hand, x could have been studying for an exam the next day, lost utility is likely to be large. In addition, x may see many different things he could be doing, for example, going out to the pub, studying, doing the garden, and so forth. To determine lost utility, he would take the mean of all the lost utilities for these activities. The reason we moderate this using trust is as discussed above: there may be something to gain from trying to establish a trusting relationship. Note that should this formula give a negative answer, this would mean that the trust x had in y was negative in the first place (unless $\widehat{U_x(\Lambda)}$ was negative, in which case x should probably cooperate anyway!), so that x would probably choose not to cooperate with y.

There are, however, situations where it might be worthwhile to cooperate even if trust is very low, for example in a situation where trust is being built up, since risks must be taken to build trust (Rempel *et al.*, 1985; Rempel & Holmes, 1986; Boon & Holmes, 1991). In these situations, the lost utility could be taken as double the negative answer, made positive. We double the result to signify the need for consequently more compensation in order to initiate cooperation with an agent we deeply distrust, or trust very little.

Having determined lost utility, x can 'name a price' which will make up for that loss. For any agent, we can put a price on each point of utility lost, say £10 for each 0.1 point of utility, thus the price for a lost utility of 0.67 would be £67.

Such a view of payment is simplistic. It serves to illustrate the ease with which a value can be worked out for any agent in a given situation. It is straightforward and easily incorporated into the formalism. It is thus simple to include in an artificial agent. As a result, it is an adequate solution to the problem in the short term, providing stimuli for future work.

5.3 Trust and the Law

At several points in this work, we have mentioned that the law can help where cooperation breaks down, or can alleviate the considerable considerations involved in determining whether or not to cooperate with other agents. The previous chapter discussed one form of quasi-legal consideration, in the form of membership of professional institutions, and acceptance of their respective codes of ethics (section 4.7.5). When an agent is known to be a member of such an institution, we can legitimately expect satisfactory work or behaviour from that agent on pain of sanctions from his institution. Above, we presented the concept of social sanctions (see table 5.1 and the associated discussion). Social sanctions must form a stronger control on agents behaviour than do professional ethics in the long run.[5] This is because society is a much larger entity than any institution, and societal sanctions can carry a lot of weight. A step further is to put sanctions into a formal framework and give them the force of law (Gambetta, 1990a). The remainder of this section presents a brief consideration of this. It is extended and commented on in chapter 8, which provides a deeper consideration based on work presented in Minsky (1991a,1991b).

A major consideration of legal systems is that things are made formal (Kowalski, 1993). In other words, for cooperative situations, for example, contracts are entered into which must be fulfilled or forfeited. It makes little difference whether either party is entering into the contract deceitfully: "Even a deceitful contract is a contract. It can, among other things, be enforced by law." (Lagenspetz, 1992, page 9). What we are concerned with here is providing a consideration of such formality. It is worth noting that, although it may seem that this legality removes the need to consider trust at all, this is not the case. It is not in every situation that we would wish to resort to legal formality in order to cooperate — this could foster mistrust and misunderstanding between good friends, for example, who would expect trusting behaviour (Deutsch, 1973; Boon & Holmes, 1991). Trust, then, works both ways, since those who are our friends expect to be trusted, and lose faith in us when they are not (Rempel *et al.*, 1985; Rempel & Holmes, 1986; Boon & Holmes, 1991). Some situations between friends are too 'hefty' to be left to trust alone, however (for example, if my friend were a furniture removal man, I may still put a bid out to contract because the obligation is too big to leave to trust, and I would not wish to offend my friend). Ongoing trusting relationships, or friendships, are fragile entities. Likewise, some situations are too small to consider legal agreements within. Moving one piece of furniture to the door is a situation in which resorting to the law would be petty and nonsensical. So, despite the truster considering a situation to be of vital importance, it is perhaps society's (or at least the legal society's) view of the importance of the situation that matters in these cases. Where money is concerned, society generally treats a situation

[5]Although this depends on the institutions concerned — teachers in Scotland, for example, can be barred from working by their institution, the General Teaching Council, of which they must be members.

as important, whereas some agents may not. Where life and limb are concerned, both society and individuals may count the situation important. There are many other considerations too numerous to mention here. See section 8.5 for further discussion. It is nevertheless the case that considerations of trust are not made redundant when the law is present. If anything they may be more so, in order to determine when to resort to the law.

5.4 Summary

Trust allows agents to consider cooperation. Indeed, trust is a necessary part of the cooperative cycle (Golembiewski & McConkie, 1975; Thimbleby *et al.*, 1994). The formalism introduced in the previous chapter and extended here provides agents with a powerful means of reasoning with and about trust. The trust, however, was always there in DAI systems: it was implicitly coded into the actions and considerations agents made. Occasionally, it was mentioned as a prerequisite for agents (Rosenschein, 1985; von Martial, 1992), but very rarely in more than a casual manner. What the formalism does is give agents the capability of using trust *in addition* to other decision making tools they have, or at least as a central tool for the evaluation of interactions.

The trust formalism, however, allows us to evaluate situations from outside. We can look into these situations, ascribe trust values and dispositions to the agents therein, and observe the behaviour of the agents, attempting to justify or explain it using trust and associated concepts. This chapter presented examples of using the formalism, and on the way, extended the formalism to take into account various short-comings or misconceptions which were highlighted in the examples. It also identified several areas where further work will be useful.

The next chapter introduces some simple principles and rules that we believe trust obeys. These are given using the formalism, which allows precise reasoning about trust to be carried through and explained. Whilst the principles have a certain intuitiveness about them, they remain to be proved or disproved. Chapter 7 gives examples of simple implementations and experiments, some of which attempt to justify and prove a number of the principles in a practical fashion.

Chapter 6

Principles for Trust

6.1 Discussion

The previous two chapters provided a definition of and extensions to a possible formalism for describing trust. The formalism allows precise discussion of the phenomenon in addition to giving us the means of implementing trusting agents in DAI. In chapter 5, we presented some examples aimed at showing how the formalism could act as an analysis and a reasoning tool for DAI and for 'everyday' life. In fact, it provides an extremely useful tool for many of the 'social' sciences, for example, sociology, social psychology, DAI, and Artificial Life. All are concerned with interactions between things, albeit at different levels of complexity. This chapter provides a different viewpoint for the formalism: that it can allow us to discuss trust *itself*, some of the principles which trust adheres to, and the benefits that trusting agents can attain over non-trusting and untrustworthy agents. The discussion ranges over evolutionary benefits, to the simple benefits to be gained from the sharing of information in a society. The principles provide a framework for working with the formulæ which will ensure that artificial trust will behave in a similar way to actual trust, resulting in trusting behaviour in an artificial agent which closely approximates real trusting behaviour. With such an simulation of trust, we may well find out some things we didn't know about the actual phenomenon in the first place (Simon, 1981).

The sample principles suggested in this chapter refer to trust in general, but it it instructive and an exercise in using the formalism to depict them using the formal language of the formalism as well as with informal descriptive language. Therefore, where, for example, $T_x(y)$ is found in the principles, we refer not only to the way in which we would want artificial trust to behave but also to the way in which trust possibly behaves in general. The principles are not, then, *rules* for the behaviour of trust, but are observations about trust from the sociological and psychological literature. Most are sensible enough to be incorporated as rules in artificial trust.

6.2 Observations

On first consideration, many of the principles and observations given below seem intuitively obvious. Indeed, many of them are, and should be, since we reason with trust all the time as humans (Luhmann, 1979), thus we all have an intimate subjective knowledge of trust. That they have not been put forward in such a manner before is understandable — one of the benefits of AI in general, and DAI in particular, is that it allows us to remove extraneous considerations from the issue at heart, and thus to consider only that which is of importance to us. The principles below, then, benefit from an insight different to those in the social sciences simply because we can give our agents trusting capabilities unencumbered by other human traits, such as happiness, fear, or anger. This allows us to build upon what we have and to include these traits as and when necessary.

6.3 Example Principles and Rules

6.3.1 Trust is Self-Reinforcing

"Trust, once established in some degree, is often self-reinforcing because individuals have stronger tendencies to confirm their prior beliefs than to disprove them." (Hinde & Groebel, 1991a, Page 187). The converse is true, as below a certain trust, individuals tend to confirm their suspicions of others (Golembiewski & McConkie, 1975). This gives us a rule of two parts, based on Golembiewski and McConkie's work, which attempts to correlate trusting and risking as self-heightening spirals. The first part of the rule is based on the idea that, if trust between two agents is initially above some threshold value (perhaps 0.00, but perhaps some other value, dependent on the agents concerned), which we call ω_x for agent x (it is therefore agent-subjective), then the trust between those two agents will not decrease below that threshold. This is because, for each member of the interaction, they will tend to look for the best in the behaviour of the other, and will tend to get it, since above the threshold, each will tend to cooperate with the other. Thus trust will increase, or at least stay constant.

The converse is true, since, if both agents trust each other below a certain threshold value, Ω ($\omega \neq \Omega$), will tend not to cooperate with each other whatever the situation, thus reinforcing the other's opinion about them as non-cooperative and unhelpful. Trust, then, is self-reinforcing. The only way out of this trap is to take a risk and co-operate (Swinth, 1967; Golembiewski & McConkie, 1975), but the further and further down the value for trust goes, the harder it becomes to justify taking that risk.

We can represent this rule using the formalism. The first part is that trust is self-heightening. Consider two agents, x and y.
If:

$$(T_x(y)^t > \omega_x) \wedge (T_y(x)^t > \omega_y)$$

Then:

$$(T_x(y)^{t+a} \geq T_x(y)^t) \wedge (T_y(x)^{t+a} \geq T_y(x)^t)$$

Informally, this says that if x trusts y more than a threshold value, and y trusts x more than a threshold value, then the amount of trust they have in each other at a later time will be greater than or equal to the amount of trust they have in each other at this time. The second part of the rules is that trust is self-reinforcing downwards also.
If:

$$(T_x(y)^t < \Omega_x) \wedge (T_y(x)^t < \Omega_y)$$

Then:

$$(T_x(y)^{t+a} \leq T_x(y)^t) \wedge (T_y(x)^{t+a} \leq T_y(x)^t)$$

This is the converse of that given above. Note that ω and Ω are not necessarily the same value for each agent, neither is ω_x necessarily equal to ω_y. Assuming a pessimistic point of view for a moment, we could suggest that in the second case, the following would happen:

$$(T_x(y)^{t+a} < T_x(y)^t) \wedge (T_y(x)^{t+a} < T_y(x)^t)$$

In other words, the trust would inevitably decrease. This is, however, unnecessary and constrains the modification formulæ given in section 4.9.2 too much. We leave the final values to those formulæ. It is sufficient to say that, when both agents do not trust each other enough, they are unlikely to at any time. A final note: for any agent x to be rational, $\Omega_x \leq \omega_x$, not the other way around. This notation should ideally contain consideration of the agent concerned, since the thresholds may well be different for different agents (i.e., perhaps we should use $\omega_x(y)$ instead).

The above considerations beg a question as to how trust behaves when the value for agents is within (Ω, ω), i.e., when the trust value is not above or below the respective threshold values for self-reinforcement. We suggest that such a state is not unlike starting points for relationships, and for interactions with agents not known, or not known too well (see Boon and Holmes (1991) for a discussion of the former). In the former, the trust will tend to build up, for example as with romantic love. To build trust, risks have to be taken (Swinth, 1967). Thus, the agent who is more disposed to take risks is more likely to build trusting relationships (Swinth, 1967; Deutsch, 1962). In the formalism, this amounts to the agent who has a higher situational trust in others, since this tends to offset high risks in the cooperation threshold. In order to build trust, then, we have to trust, both our initial judgements, and other agents. An example may help. Consider x and y who have never met ($\neg K_x(y)$ and $\neg K_y(x)$). If x is trusting, he may assign a value equal to T_x, which is likely to be high in any case (x is trusting) to the measure $T_x(y)$. This being the case, it is likely that the resultant situational trust will exceed the cooperation threshold. Take the example for y considering x in the previous chapter, for instance. If this is so, x will cooperate with y, and y will increase trust in x, and so on. Conversely, if T_x is very low, the chances of cooperation are lower, thus the chances are that the resultant trust between the two will drop rather than rise. This simple example may help to make the situation more clear. In between values of Ω and ω then, anything can happen.

6.3.2 An Increase in Trust Increases Societal Knowledge

Considering a society where information is of value, such as our own, the societal knowledge can be characterised by the amount of knowledge that the members of the society know collectively (i.e., that the majority of members of that society know). It follows that, if the only way information is disseminated is for agents to share it, that when agents trust each other more they will share more information, thus the amount of disseminated knowledge in the society will increase. The converse does not hold — if suddenly all agents stopped sharing information, the amount of societal knowledge would not decrease as suddenly, rather it would stay static, perhaps stagnating. It would not, however, increase. See Bok (1978) for a discussion. Briefly, the following may hold:

$$K(\mathbf{S}) = \sum_{a \in \mathbf{S}} K(a)$$

Thus the total knowledge in society is a summation of the knowledge of all agents in that society. If agents share information with others, this will increase, and if not, it will remain, at best, static:

$$\forall a, b \in \mathbf{S} : \text{If } T_a(b) \Rightarrow \text{Will\_share\_knowledge}_a(b)^t$$

And so:

$$K(b)^{t+1} > K(b)^t$$

In other words, conditional on b not knowing the information a told her before she was told, her knowledge increases, and so does that of society. This is a simple exposition of knowledge, and does not take into account the large amount of literature on the subject. Nevertheless, it does show the worth of the formalism in describing such occurrences.

6.3.3 Dissemination of Trust Knowledge

What happens if three agents are in a situation where one of them knows only one of the others, but wants to consider the third? Formally, the situation can be described as follows, so as to avoid ambiguity. We consider agents x, y, and z. Agent x knows both y and z: $K_x(y) \wedge K_x(z)$, the same applies to y: $K_y(x) \wedge K_y(z)$. However, z knows only y: $\neg K_z(x) \wedge K_z(y)$ holds. The situation involves z having to consider x. Because z knows (and by definition has a trust value for) y, z can ask y for help in this matter (and assuming y gives it and z does ask for it). The rule we advance here concerns how much z will trust x, the value of which will be dependent on how much z trusts y, and how much y trusts x. The following rule will hold if z is rational:

$$\begin{aligned}&\neg K_z(x)^t \wedge K_z(y)^t \wedge K_y(x)^t \rightarrow \\ &\quad K_z(x)^{t+a} \wedge (T_z(x)^{t+a} \leq T_y(x)^t) \wedge (T_z(x)^{t+a} \leq T_z(y)^t)\end{aligned}$$

Thus, z will not trust x any more than y did, but also, because the amount of trust z will have in x is mediated by the trust z has in y, the information giver, the resultant trust will be no greater than the amount z trusts y in the first place. A simple formula for determining trust in this kind of example is as follows:

$$T_z(x) = T_y(x) \times T_z(y) \tag{6.1}$$

Naturally, this can be extended an arbitrary number of times, forming a kind of 'trust network.' This brings about questions relating to how one measure of trust compares with another for another agent. Clearly, should I say I trusted someone 50%, and my friend said the same, we may not both mean the same thing — my 50% worth may mean more to me than my friend's. This perhaps highlights one of the problems in the use of values. A discussion of why values are used can be found in chapter 2.

6.4 Transitivity

Trust is not transitive. Although assuming x is rational, the following will hold:

$$(T_x(y) > T_x(z)) \wedge (T_x(z) > T_x(w)) \Rightarrow (T_x(y) > T_x(w))$$

This is a truism for numbers, but it is possible to imagine an irrational agent who may act as if it was not true. Trust is not just a number, rather a decision made by an agent. The resultant behaviour would be odd, but nonetheless possible for trust.

It is not the case, however, for any relation '>':

$$T_x(y) > T_y(w) \not\Rightarrow T_x(y) > T_x(w)$$

In other words, that x trusts y by some amount and y trusts w by some other amount says little or nothing about how much x trusts w, if at all. Indeed, it may well be the case that $\neg K_x(w)$, even if $K_y(w)$ and $K_x(y)$.

6.5 Knowledge

The above discussion raises another important point regarding knowledge. It is not necessary for a trusted party to know she is trusted, let alone know the agent who is trusting her:

$$T_x(y) \not\Rightarrow K_y(x)$$

An example of this might be based in politics: the electorate choose their leaders, but the leaders have little conception of individual members of that electorate, despite the trust placed in them by those individuals. In this example, the elected leaders may (should) know of the trust placed in them, however. An example of trust placed in someone who knows nothing of it might be the trust we place in drivers of cars in the street where we send our children to school: individual drivers know nothing of us, whether or not they see our children, yet we still trust them to drive carefully (and this is trust, since we have a right to ask this, or expect this, of them).

An interesting point arises from this:

$$T_x(y) \not\Rightarrow K_x(y)$$

In words, it is not altogether necessary for x to know y before trusting him. The case of the car driver is illustrative. This assertion is a little more radical, and a little more limited, than most of the others in this chapter. Clearly, something must be known, or at least, assumed, about the trusted party in order for us to be able to trust them. In the instance of the car driver, we assume he has passed his road test and is not drunk, and is not under the influence of drugs, and so forth. Indeed, we assume many things about those whom we do not know in the proper sense of the word. It is, clearly, enough to trust them. Indeed, this is an example of the trust we place in society to reduce the complexity of everyday life (Luhmann, 1979). To seek to know everything about those we trust would soon result in cognitive overload, no less for artificial agents than for humans. It is fair to assume a limited knowledge where trust is concerned, then. This could be represented by $\widehat{K_x(y)}$ in the formalism.

6.6 Rational Trust

Many of the principles and formulæ presented thus far in the thesis discuss rationality in greater or lesser detail. Since we would wish to assume that our agents are rational (Simon, 1955; Simon, 1981; Gladstone, 1961), these have suggested what rules rational trusting should follow. We collect and expand them here.

6.6.1 Ordering relationships

The rational agent will have trust values satisfying the following relationships:

$$(T_x(y) > T_x(z)) \land (T_x(z) > T_x(w)) \Rightarrow (T_x(y) > T_x(w))$$

The discussion surrounding this assertion is given above. It should be clear enough to require little justification.

6.6.2 Minimum and Maximum Thresholds

In the discussion above pertaining to spiraling trust (both upwards and downwards), we postulated the existence of threshold values above which or below which trust spirals. The threshold above which trust spirals upwards is notated ω_x for agent x. The threshold below which trust spirals downwards is notated Ω_x for the same agent. For x to be rational, the following must hold:

$$\omega_x > \Omega_x$$

The justification for this is similarly straightforward.

6.6.3 Increases and Decreases in Trust

When considering a simple trusting agent (i.e., one who does not use rules of reciprocation — see chapter 4), the trust he has in a trustee will ordinarily increase if cooperation occurs, and decrease otherwise. The amount of the increase or decrease depends on the costs or benefits of the situation that have been incurred. A simple rule for the adjustment of the general trust of agent x in agent y ($T_x(y)$) is:

$$\text{Cooperates}(y)^a \Rightarrow T_x(y)^{a+1} > T_x(y)^a$$

And:

$$\text{Defects}(y)^a \Rightarrow T_x(y)^{a+1} < T_x(y)^a$$

Making this more complex, in order to take into account other considerations, most importantly reciprocity, is not difficult. The rules for this are given in chapter 4, section 4.9, but for completeness we repeat them here:
If:

$$\text{Helped}(x, y, \alpha)^{t-\delta} \wedge \text{Defected}(y, \beta)^t \tag{6.2}$$

Then:

$$T_x(y)^{t+1} \ll T_x(y)^t$$

The converse is if:

$$\text{Helped}(x, y, \alpha)^{t-\delta} \wedge \text{Cooperated}(y, \beta)^t \tag{6.3}$$

Then:

$$T_x(y)^{t+1} \geq T_x(y)^t$$

Taking into account costs and benefits is likewise not a difficult task. If the costs are high following a defection by the other agent, trust will decrease by a larger amount that if the costs had been lower; thus we reduce trust by some variable, $\mathcal{C}$, which is dependent on the costs incurred (and could conceivably take other things, such as reciprocity or moralistic aggression into account (Trivers, 1985). We then have:

$$\text{Defects}(y, \alpha)^t \Rightarrow T_x(y)^{t+1} = T_x(y)^t - \mathcal{C}$$

The same could be suggested for benefits following cooperation (for which we increase trust by a value dependent on the benefits gained, $\mathcal{B}$):

$$\text{Cooperates}(y, \alpha)^t \Rightarrow T_x(y)^{t+1} = T_x(y)^t + \mathcal{B}$$

6.7 Evolutionary Ideals

6.7.1 Reciprocation, Evolution, Trust

In a society of agents, interactions between agents will proceed apace. Indeed, it is likely that two agents will meet more than once, if the society is small enough, or if the two agents are working in the same area, geographically or otherwise. In such interactions, it is an evolutionary strength to be a reciprocal cooperator (Harcourt, 1991). In other words, an agent, a, will do better if he reciprocates help with another, b, than if he did not. Reciprocation is a common form of behaviour in the animal world (Harcourt, 1991; Trivers, 1985). The great apes help each other in fights, for example, in the hope (or knowledge) that their status is increased in the troupe, and that help will come their way from those they help (Trivers, 1985). Vampire bats feed those who have not eaten well one night, in the knowledge that they in turn will be fed on a bad night (Harcourt, 1991). It follows that, if the bats who fed others were not in their turn fed, they would die out, and so would their altruistic behaviour. They are fed, however, because this apparent altruism enables the population as a whole to survive, and ensures the survival of individual bats.

Reciprocal altruism is a form of trust in the animal world (Harcourt, 1991). In more intelligent animals capable of reason, reciprocation is more common if some form of trust as we understand it is present. If trust was not present, the initial risks would not be taken to be altruistic (Rempel *et al.*, 1985; Boon & Holmes, 1991), reciprocation would not occur, and the collective society would be worse off. To illustrate this, consider the following example.

If two agents, a and b were reciprocating trusters, and another two, c and d, were not disposed to cooperation at all, and if they were members of a team which hunted by night, the following situation may hold. On any one night, there will be the chance that any of the agents may not get enough food. In terms of payoffs, this situation is given the value 0. If an agent gets enough food for itself, its payoff is 3, and if it shares this food, the payoff to each agent sharing the food is 2. We consider two consecutive nights. On the first, a does well, b does not, and a shares its catch. Payoff to each = 2. On the same night, c does well, d does not, but c refuses to share its food, thus payoff to c is 3, to d, 0. On the next night, b does well whilst a does not, and so, since b is a reciprocator, it shares its food with a. The cumulative payoff for each is 4. Likewise, d does well, c does not, and so the cumulative payoff for each of these two, since d refuses to share its food, is 3. So despite the short term gains c and d make by not reciprocating or trusting, the long term result is a fitter a and b, and this fitness can translate into leadership of the troupe, for example. Reciprocation can pay off in such a circumstance.

The above example is somewhat contrived, but serves to illustrate the point that long term reciprocation is good for individuals and the group as a whole (Harcourt, 1991; Trivers, 1985). Of course, on a good night, all agents get a payoff of 3. There will be at some time, however, a bad night for some of the agents, and it is then that the differences arise. A run of bad nights for d will result in its death from starvation, and a famine will result in the deaths of non-reciprocators, whilst those who reciprocate and are altruistic may survive.

This general principle can be made use of in the formalism, particularly with regard to the alteration of trust values. See the previous chapters, also chapter 8, for further discussion. In summary, reciprocation is more likely in a trusting relationship (Boon & Holmes, 1991) and reciprocation is good in evolutionary terms (Harcourt, 1991).

6.8 Summary

This chapter presented and discussed several 'principles' which trust, or at least the formalism for trust presented in the previous chapter, appears to follow. In some senses, the principles themselves are of limited importance. What is important is that the formalism provides the means to discuss such principles precisely and clearly, and with much less ambiguity than has previously been the case.

That trust can be discussed with little ambiguity is one of the goals of the thesis. Another is that the formalism can be implemented and embedded in an artificial reasoning agent. The following chapter presents discussions and implementations of the formalism, with the aim of proving the utility of both the formalism and trust *per se* for such agents.

Chapter 7

Practical Work

7.1 Introduction

There have been several investigations into the concept of trust in the past. They ranged from experiments involving questionnaires (Rotter, 1967; Rempel & Holmes, 1986) to more abstract visions based on philosophy, sociology, social psychology, and other social sciences (Barber, 1983). None of them provided a usable formalism for trust, and very few attempted a workable explanation of the concepts involved. As far as is known, none have thus far provided an implementation of their ideas.

The present work goes beyond other explorations of trust by providing two major contributions:

1. A workable formalisation;

2. A simple implementation to demonstrate the concept.

The first of these was discussed in earlier chapters. The remainder of this chapter concentrates on the second, but is limited in its realisation. It is, however, an indication of what is possible, and an important confirmation that the formalism introduced here is workable and can be embedded in artificial agents.

The experiments described below were in fact carried out several times, but the amount of data generated by each was quite large. It was decided to concentrate on and report *typical* examples of each experiment with the aim of showing the general style of results for each.

7.2 Implementations and Experiments

There are two major implementations that have been carried through for this work. Both are concerned with the Iterated Prisoner's Dilemma (PD) (Axelrod, 1984; Behr, 1981; Dolbear & Lave, 1967). There are reasons for this choice.

1. The PD is a well-known, well-understood 'game,' in the social sciences and in computing. Consequently, any results or insights gained in the experiments carried out here will have a wide audience and applicability.

2. Being constrained, the PD provides us with the ideal tool for experimenting with agents since, from the agent's point of view, the only choices available to them or the agent with whom they are interacting are cooperation or defection. Such constraints allow the agent to reason in a limited fashion, whilst still providing us with some measure of the result of such reasoning.

3. Because the PD is so limited, it follows that simple trusting agents can easily estimate costs, benefits, and utility for a situation for themselves and also for those with whom they interact. In the future in more complex interactions, such considerations will be agent-subjective and estimated by the agents themselves.

4. Payoffs in the PD present an ideal measure of how well particular agents, and indeed the society as a whole, are performing. This measure can be used to ascertain the relative successes of particular strategies as a whole or in interactions with other strategies (or themselves), and could be used in a form of 'survival of the fittest' to show the relative evolutionary stability of particular strategies (Lomborg, 1992).

In short, the PD provides us with a seemingly ideal tool for performing research and fine tuning trust strategies. There are drawbacks to its use, however:

1. The PD is by its very nature confrontational. This must be the case in its classic form, otherwise no dilemma exists, because "noncooperative short-run maximising behaviour is inconsistent with long-run (cooperative) behaviour" (Dolbear & Lave, 1967). Agents must naturally enter a PD interaction in a confrontational state of mind — hence the dilemma. For experiments in trust as we see it, this is an inherent drawback since we see trust as an avenue for

achieving cooperation where it may not seem possible. The PD makes this more difficult.

2. Because of the nature of the PD, in order to 'do well,' i.e., score more points, or increase fitness, one has to defect before one's opponent (Behr, 1981). This is unfortunate, but since it is a phenomenon that is well known, can be taken into account when performing experiments.

3. Objections to the PD have been raised in other areas, notably by Argyle (1991), who lists the following differences between the PD and everyday life:

    (a) *Play is simultaneous, with ignorance of the move of the other player, and there is a risk inherent in the other's failing to cooperate.* Argyle suggests that in real life, if a person decides not to do something, their payoff is 0. Thus, if A decides not to play tennis or join in house building, his payoff is again 0. Whilst this is not strictly the case, we leave this argument for the present.

    (b) *The game is too abstract.* In this instance, this is of little importance.

    (c) *There is usually no communication.* In situations where communication takes place, cooperation is generally increased (Argyle, 1991; Deutsch, 1962). Since the PD generally disallows communication, this may hinder its results. In the case of the artificial agents used here, this is not too much of a problem, since we do not expect communication to produce bargaining behaviour at present.

    (d) *The players are usually strangers, and are invisible to one another.* This is not always the case in the following implementation.

    (e) *Social norms are absent.* Cooperative or defective norms are absent since the PD, being abstract, is a game without such rules. In the case of the following implementation, again, this has little significance since social norms are learned from childhood (Argyle, 1991), and our agents exist 'as is,' with no remembered 'childhood,' although such a consideration could be made in the future.

Despite these objections, the PD provides a constrained, simple, yet realistic mirror of the kind of interactions that we would expect trusting agents (or any artificial agent) to be involved in. Such interactions are likely to be characterised by many of the same things that the PD is characterised by, such as:

1. Each interaction has specific ends (to increase payoff, to gain information, to enlist help) and agents are aware of these ends (or can at least make an 'educated' guess about them.

2. The respective payoffs are readily known, or can be estimated.

3. The other agent may or may not be known. If known, then memory can help in the decision making process.

4. The result of an interaction is generally clear cut (cooperate or not, give information or withhold it, give or withhold help), although for artificial agents this is less true in many situations — the giving of information, for example, can be partial, or somewhat clouded.

7.3 Tournament — Robert Axelrod and Peter Pang

The first of the experiments was modelled on Axelrod's PD tournament (Axelrod, 1984). In his book, Axelrod omitted the algorithms for the strategies that were used, however. This oversight was corrected by Pang (1990), who repeated Axelrod's experiments, and provided some extra insights into the results. In addition, Pang provided algorithms and code for the strategies and the tournament. These algorithms and code were used as the basis for the first implementation of trusting agents.

The first implementation was in fact particularly simple and experimental — it was carried out primarily to ascertain if the formalism could be implemented, and whether it would work as expected or not. In addition, the algorithms used for the trusting agents were continuously in flux, and do not match precisely what has been presented as the formalism in chapters 4 and 5. Since this is the case, we mention the findings of the tournament only briefly here.

The tournament was as discussed in Axelrod (1990), with the addition of simple trusting strategies, one a simple truster who used trust estimates all the time, one who trusted 100% until a defection occurred, then resorted to trust estimates, starting at 50% (called *trust-for-tat*), and one who was a trusting agent, but had a specific propensity to trust (of 50%), no matter what its trust value was (called *propensity-trust*). Each run of the tournament had each agent pitted against each other agent (including themselves) 200 times.

The results of the tournament confirmed two things:

- *The formalism can be implemented.* This was a major result in terms of the present thesis.
- *The trust strategies behaved roughly as expected.* Since the strategies were probabilistic, and relatively simple, this was not a particularly illuminating result.

The results of the tournament are of little practical applicability here because of the simplicity and limitations of the strategies proposed. In addition, the following implementation carried many of the hallmarks of the tournament, but placed them into a social and spatial sphere, which, from the point of view of trust, is somewhat more important.

7.4 Society

Whilst the tournament experiment provided some interesting insights into trusting behaviour and how well it fared against other strategies, it was somewhat artificial, for the following reasons:

1. *Each agent met each other agent for a specified number of interactions.* Although the number of interactions was unknown to the agents, this is still a limitation, and does not accurately mirror 'real world' situations as well as we would wish.
2. *The payoffs were the same in each interaction.* There was no concept of 'different' situations for agents. This is a less important objection than the first, but nonetheless does not test the particular aspects of trust that we would want.
3. *Nice Guys Finish Last.* Behr pointed out, as mentioned above, that doing well in such a tournament involves defecting before one's opponent (Behr, 1981). This is hardly an ideal situation, and cannot be avoided if we wish the same number of interactions between each and every agent.

It should be clear that the tournament was not the ideal solution to finding an experiment using the PD which tested the particular strengths and weaknesses of trust as a strategy for decision making. It was decided that what was needed was a more versatile arrangement which retained all of the benefits of the PD presented above yet provided the following:

- a graphical interface which can show the movements of individual agents.
- freedom for agents to 'move around' in an artificial environment.
- a range of different situations, with possible different payoffs and payoff structures for each.[1]
- the concept of a society of agents.
- a limited form of evolution for strategies.
- the ability to add or delete new agents, strategies for agents, or situations (payoff structures, etc.) at any time.

Later we added:

- freedom for agents to choose with whom to interact in society.

In fact, the final addition is less desirable that it might at first seem — we do not wish our agents to develop insular cliques which do not interact with the rest of society, since that perpetuates the problems discussed elsewhere in this thesis of lack of information distribution, lack of cooperation where it might otherwise be possible, and so forth. It is also less obvious that this arrangement mirrors the world at all closely, at least not the world artificial agents may operate in. Unquestionably, they will operate in diverse and unknown situations, probably with unknown agents a lot of the time. Care will be needed to avoid exploitation, yet some interactions will be necessary (take the movement of furniture, for example — one agent cannot move a settee alone, yet two can, and so cooperation is necessary to move the settee). At least if a choice is available, agents will tend to lean towards interactions with previously known, trusted parties rather than with unknown or untrusted parties. Two such possible leaning structures are shown in table 7.1. The second column of the table suggests an agent with the attitude 'better the devil you know,' the first that the

[1]Whilst still keeping the outward PD form, i.e., Cooperate or Defect, and likewise for the other agent, with decisions made before other's decision is known.

Rank	Possible Leaning	Possible Leaning
1.	Known and trusted.	Known and trusted.
2.	Known and impartial.	Known and impartial.
3.	Unknown.	Known and not low trust/not trusted.
4.	Known and low trust/not trusted.	Unknown.

Table 7.1: Possible leanings towards an agent (strongest first)

unknown can be no worse than what is already experienced (if it is bad). There will be other configurations.

In any case, the final arrangement was limited so that interactions between agents were forced if they met (a forced PD between agents which 'bump into' each other), and that agents can only 'see' a short distance, and are thus limited in their ability to move away from untrusted agents and toward trusted. In addition, there is a random factor in all movements which may result in agents being sent in directions they may not wish to go.

The advantages of these arrangements are clear:

1. Interactions are no longer of a prescribed length and can be terminated at any time. Thus, if an interaction is going particularly badly, there is an 'escape' option. The reverse is also true for interactions which are going well.

2. The concept of a group can form as a coincident of this arrangement.

3. We introduce the concept of space — agents move around, not staying in one place, and interact with agents they meet. Thus:

   (a) Agents are more likely to interact with some agents rather than others.

   (b) As an extension of this, agents are more likely to interact with others *in their vicinity or neighbourhood*, again as could be expected in a real world implementation.

Unfortunately, evolution in such a space is fairly indiscriminate — those agents who have had the misfortune to be geographically isolated are treated as badly (if not worse) than those who have had many interactions but have lost out in most of them.

7.4.1 Details of Implementation

The implementation was coded in HyperCard on an Apple Macintosh (Claris, 1990). From the outset, the design decisions outlined above were adhered to in that the system had to be made as flexible as possible, allowing for the addition and deletion of new agents, strategies and situations (payoff structures). The practical details of the implementation are as follows:

Main Interface

The implementation is based on the concept of a small testbed, called the 'PlayGround,' within which agents have total freedom of movement.[2] The PlayGround is a simple 10 by 19 square surface, with one square able to be occupied by one agent at any time. A snapshot of the PlayGround is shown in figure 7.1.

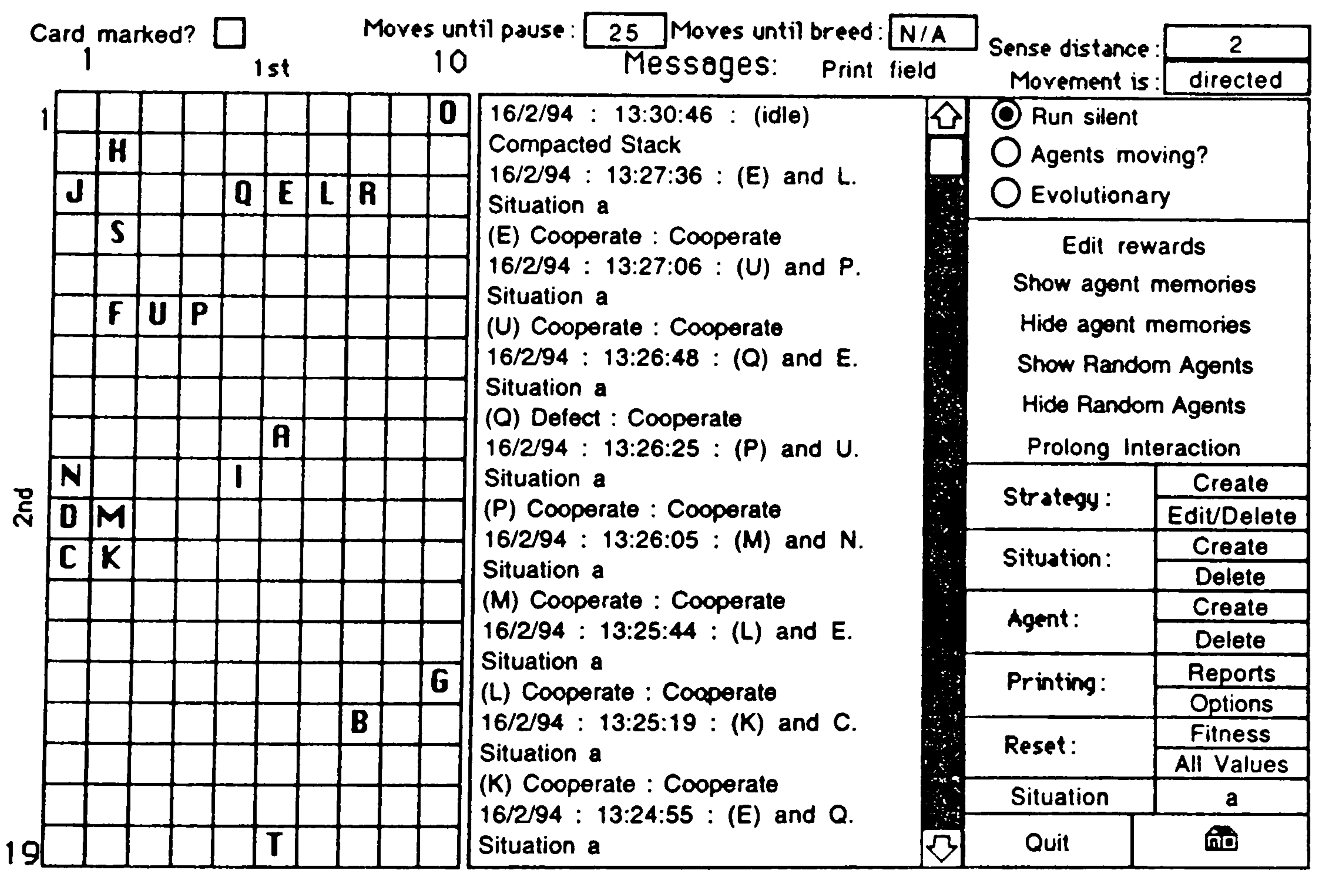

Figure 7.1: The PlayGround: A simple testbed for trusting agents

The actual PlayGround is shown to the left of figure 7.1, and some controls are on the right. There are several options which can be exercised using the controls in

[2]This is not the same as total freedom of *choice* of movement, which agents do not have. See later for agents with limited freedom of choice of movement.

order to make movement random or directed (see below), whether to turn 'evolution' on or off, to create new strategies for PlayGround agents, new payoff structures for interactions ('situations'), and to print or reset specific values.

Movement for agents was initially random, although this was later changed to reflect the comments above, and agents were allowed to migrate towards more trusted or known agents, in a limited fashion. Range of 'vision' for agents is limited, and can be set from the main interface. For the purpose of most of the following experiments, it was set to 2 only, i.e., agents can see 2 squares north, south, east or west. If an agent 'sees' a trusted other within that range, then a weighting is put on movement in that direction, but the final result is still based on a random number generated by the machine. This is to reflect the idea that, ultimately, we might want our agents to perform tasks which involves them putting themselves into places they might not wish to go.

Interaction is achieved by a forced Prisoners' Dilemma whenever an agent attempts to move into a square which is already occupied. Both agents in an interaction are considered equal, although the agent that is originally on the square can be assumed to be the agent who is 'asked for help,' or 'doing the trusting' in the situation (in fact, both agents can consider using trust or whatever strategy is set up for them). For any one agent, there is a strategy which that agent uses. This can be changed easily. In addition, trusting agents have dispositions, as discussed in chapter 4. For the purposes of the following experiments, we used three dispositions: Optimist, Pessimist, and Realist. Each disposition followed a slightly different rule in ascertaining how to trust. The rules were as suggested in chapter 4. Briefly, for estimating general trust, they are:

- *Optimist:* Take the highest trust value available (remembered) for this particular agent in this situation. If not known in this situation, take the highest trust value of any situation it is known in. If not known at all, take the most recent basic trust, and put it into general trust.

- *Pessimist:* Take the lowest trust value available (remembered) for this particular agent in this situation. If not known in this situation, take the lowest trust value of any situation it is known in. If not known at all, take the most recent basic trust, and put it into general trust.

- *Realist:* Take the mean trust value for all past interactions with this agent in this situation. If not known in this situation, take the mean for all other situations. If not known at all, take the most recent basic trust.

Alterations to trust were not affected by the trusting agent's disposition (see below for a discussion of this).

In any interaction, a random situation type of the list of situations known is chosen, and both agents are informed of the situation name; it is common knowledge that both agents know the situation name and the payoff structure for this situation. There is the possibility of setting a 'global situation type' that all interactions subsequently take place in. This allows a more accurate view of the relative fitness of different strategies (see later). After both agents have made their decisions, fitness values are updated according to the payoff structure for the particular situation. For example, if the payoff matrix for two agents, A and B was as follows (see appendix A):

		B *c*	**B** *d*
A	*c*	3, 3	0, 5
A	*d*	5, 0	1, 1

And both cooperated (*cc*), then each agent's fitness would be increased by 3.

Following each interaction, if the agent is a trusting agent, it adjusts its trust values. Generally speaking, if the other cooperates, trust is increased, and if the other defects, trust is decreased. For the purposes of the experiment, it was decided that the amount of increase or decrease be standardised, to allow a more meaningful examination of the trust concept. There are 4 possibilities from the point of view of any one agent in an interaction. The alteration of the trust values from the point of view of **A** interacting with **B** are as follows:

1. **Both cooperate.**
   Basic Trust increased by 1%
   General Trust in other increased by 10%

2. **A cooperates, B defects.**
   Basic Trust decreased by 1%
   General Trust in other decreased by 10%

3. **A defects, B cooperates.**

   Basic Trust increased by 5%

   General Trust in other increased by 1%

4. **Both defect.**

   Basic Trust decreased by 5%

   General Trust in other decreased by 10%

The reasoning behind this set of adjustments is described in detail in chapter 4, and discussed further later in this chapter.

Following adjustments, the cycle begins again, with agents moving. The full cycle, then, is:

```
repeat
 for each agent do
  begin
     move
     if another agent 'met' then
         interact
         adjust trust values if necessary
         save results
      end if
   end
until maximum number of moves reached
```

The maximum number of moves referred to is set by the user at the start of the experiment. It can take any value within practical limits (in practice, the maximum number of runs should be kept lower than 400 – 500, since it was found that the time taken for interactions increases steadily as time advances. This is a problem with the implementation platform, HyperCard. With different platforms, the speed would increase markedly).

7.4.2 Experiments, Results

Several experiments were run with the PlayGround. Initial simple experiments are detailed below. More complex experiments and results are discussed later in the chapter, and some results are summarised in appendices A and B.

We label the experiments $\mathcal{A}, \mathcal{B} \ldots$

Experiment $\mathcal{A}$

Experiment $\mathcal{A}$ was a set of small experiments, involving simple trusting agents. The trusting agents were of two of the three possible dispositions: Optimist and Realist (see chapter 4 for a discussion of this). The aim of this experiment was to see if trust could be educated. That is, if from a low estimate of trust, one which would result in defections initially, a trusting agent could be educated into taking the risk of trusting (actually by increasing its trust value to a level at which, and above which, cooperation would ensue). The taking of a risk is an essential part of attempting to build a trusting relationship (Lewis & Weigert, 1985), and it is useful to see how much of a risk has to be taken to prompt such trusting. There are, naturally, different answers to this, depending on which agent we are trying to educate.

This experiment was static. There were two tests, initially set up artificially to create the kind of situation we were looking for. In the first test, there were two agents, A and B, in one situation, a. The payoff matrix for the situation was as follows (also given in appendix A):

		B	
		c	d
A	c	3, 3	0, 5
	d	5, 0	1, 1

Both A and B were realist simple trusters. Initially, A was inclined to distrust B, with $T_A(B) = 0.08$, and calculations setting $T_A(B, a) = 0.39$, whilst $\text{Coop\_Thresh}_A(B, a) = 1.297$. Clearly, A defected. For B, however, the situation was somewhat different, $T_B(A) = 0.86$, so $T_B(A, a) = 3.71$! Since $\text{Coop\_Thresh}_B(A, a) = 0.054$, B was well disposed to cooperate. The test was run for 112 iterations before stopping. At the end of this, B was still inclined to trust, but by this time $T_B(A) = 0.59$, and $T_B(A, a) = 2.84$, with $\text{Coop\_Thresh}_B(A, a) = 0.09$. (Still a long way to go for a defection.) On the other hand, A was becoming more inclined to cooperate, with $T_A(B) = 0.09$ (but, notably, T_B had increased from 0.08 to 0.243), and $T_A(B, a) = 0.44$, $\text{Coop\_Thresh}_A(B, a) = 1.01$. It seems clear that A will eventually cooperate, but at some considerable cost to B — after 112 iterations, A's fitness[3] was a large 555, with B's still 0. Clearly not a good situation for B. Sometimes, risks have to be taken to develop trust (Govier, 1992; Lewis & Weigert, 1985), but B seems to be overstepping the bounds of rationality.

There are reasons for this. Firstly, the amount of trust A started with was extremely low, and thus it had to take a lot of persuasion to be encouraged to trust. In addition, since A had a potentially unbounded memory span, the very low starting point for trust simply continued to have a large effect on the final estimate for $\widehat{T_A(B)}$. There are ways around this — if A's memory span was short, cooperation would arguably have ensued much earlier, and this is tested in experiment $\mathcal{B}$, below. In addition, if the disposition of A had been optimism, cooperation would again have started earlier, with much less of a loss for B. It is this that the second test of this experiment looked at. The starting points were the same, with the payoff matrix as given above. The agents' names were changed for clarity, and A became C, B became D. One other change was made, and that was that C became an optimist according to our definitions. The initial decisions by both agents were identical to those for A and B. However, the trust C had in B continually rose without the hindrance of its low starting value, and after just 13 interactions, C chose to cooperate with D, with $T_C(D) = 0.09$, $T_C(D, a) = 0.62$, and $\text{Coop\_Thresh}_C(D, a) = 0.57$. The value for $T_C(D)$ seems low, but it was enough for cooperation to ensue, and quickly (before another 20 iterations were up), both agents' trust in each other was extremely high. Detailed output from the second part of experiment $\mathcal{A}$, part two can be found in appendix A.

What can be learnt from this?

- As was initially expected, the optimist learned cooperation much more quickly than the realist (although the difference was marked and unexpected).
- The realist, A, presents us with special problems — its trust was extremely low to start with, and due to the way we calculate trust estimates, $\widehat{T_A(B)}$ stayed low because the initial value weighed it down. This matter is further addressed in the next experiment.
- The realist with high initial trust is shown to be particularly vulnerable here.

[3]In fact, after 170 iterations, cooperation had still not endsued.

Almost it seems to be an unconditional cooperator.

- A lot of the results here are moderated by the very large utility that the situation held for the agents. Again, this is an area for future attention. At the present, the agents calculate their utilities in a very simple fashion according to the payoff structure for the situation, and occasionally, this results in an artificially high value which clouds the final result. In this situation, however, the utility for each of the agents was the same, so we can comment on their behaviour as dispositions relative to each other.

Experiment $\mathcal{B}$

Identical tests to those documented in $\mathcal{A}$ were run, but this time, the memory span of each agent was reduced from its potentially unbounded value to just 10. The agent, then, could remember only ten interactions into the past *with each other agent.* In other words, even if the last interaction with this particular agent was, say 20 interactions ago (20 interactions with other agents), it would still be remembered, as would the previous 9 interactions with this agent. The results are very different from those obtained in experiment $\mathcal{A}$.

- With A and B in the same initial configuration, cooperation ensued after 127 iterations, from which time cooperation was guaranteed, barring any severe misunderstandings.
- For C and D, things were identical, with cooperation again ensuing after just 13 iterations.

The reason for the second finding is clear: for an optimist, the highest remembered trust value is always taken, and so, since in this situation D always cooperated, C's trust always increased. Thus, the highest trust value was always the last known trust value. Reducing memory span does little in this situation. The outcome is different, however, in less cooperative circumstances, if D is random, for example, or starts to defect.

The first finding is more interesting: cooperation ensued much earlier than it would for the equivalent experiment with potentially infinite memory spans, and the reason is that, with a shorter memory span and increasing trust, the mean of the last ten interactions' trust values will also steadily increase, but *without* being held back by its initial low values.

Of course, these results can work both ways — the optimist who can only remember five interactions back in circumstances of decreasing trust will quickly lose faith, whilst the optimist who can remember all such interactions will not as quickly. Likewise, the realist who started off on a high trust will find that trust diminishing more quickly if she experiences continual defections.

Experiment $\mathcal{C}$

The third experiment went on to extend the findings of experiment $\mathcal{B}$. The memory span of both A and B was reduced to 1, whilst still keeping initial values and dispositions the same as in the previous two experiments. The result of this set of interactions was that after 22 iterations, cooperation ensued. This is drastically less than for the first part of experiment $\mathcal{A}$. It is more than was experienced in the second part of $\mathcal{A}$, however. The reason for this lies in the way the implementation calculates the mean for trust: values are stored for all situations experienced, with all agents. There are two entries for each interaction: a before entry and an after entry. The mean of all such entries is taken for the realist to calculate $\widehat{T_A(B)}$, thus, for a memory of 1, two values are actually taken and used. In reality, a memory span of 1 should result in the realist, the optimist, and the pessimist being indistinguishable. This is, then, a problem with the implementation, not the formalism.

7.4.3 More Detailed Experiments

The experiments detailed above were intended to ascertain whether the behaviour of various dispositions was as expected, along with whether trusting behaviour can be learnt, given the right prompting. Clearly it can, and the amount of risk taken, or costs absorbed, depends a great deal on who is being interacted with, and their disposition.

More detailed experiments performed with the testbed are also possible, and we document some here. They involve, amongst others, the addition of a new strategy — the random cooperator, the use of more than one payoff structure for interactions (i.e., the introduction of more than one situation), many agents in the grid, moving at random (no direction has a greater weighting than any other), agents moving with a purpose (towards or away from particular other agents), and finally evolution. Most of the following experiments were run over 250 iterations of the above loop of move, interact, update for each agent. Other variables are documented below. Results for each experiment are briefly discussed after their description. In addition, snapshots of sample runs for some of the experiments can be found in appendix B. Table 7.10 summarises the starting points for each of the more complex experiments.

The Random Cooperator

The next experiment (experiment $\mathcal{D}$) involved the introduction of a new strategy, which cooperated or defected at random. There was an equal chance of either occurrence in any interaction. In a field of 21 agents, we had 6 of them random, 5 optimist simple trusters, 5 pessimist, and 5 realist. The experiment was run for 250 loops. Memory for the trusting agents was set to a span of 10 interactions. The payoff structure for the interactions was fixed, and set to (e.g., for agents A and B):

		B	
		c	d
A	c	3 / 3	0 / 5
	d	5 / 0	1 / 1

Summary of results

A league table of the top 5 ranking agents (those agents with a fitness greater than the others after the iterations finished) is shown in table 7.2. This method of determining the most successful agent is in fact a little inaccurate, since some agents are necessarily going to interact less than some others. Ordinarily, those who interact the most get a higher fitness. This problem is partially addressed in experiment $\mathcal{G}$ and others where movement is 'directed' (see below).

In fact, C interacted with many other agents, but defected almost all the time, thereby notching up a high score as a free-rider. In a society where communication was prevalent, such actions should not be possible, since agents would be able to inform each other of C's behaviour. Other random agents did quite poorly, with E attaining a fitness of 37, D, a fitness of 55.

Ranking	Name	Disposition	Fitness
1	C	Random	230
2	I	Pessimist	202
3	A	Random	174
4	B	Random	163
5	D	Random	155
Mean			105
Minimum	G	Pessimist	18

Table 7.2: Summary of results: experiment $\mathcal{D}$.

Different Payoff Structures

The introduction of new payoff structures serves to provide different 'situations' within which agents can interact. Different payoff structures inevitably lead to different utilities, perceived costs and benefits, and thus possibly different decisions for the same agent (see chapter 4). In experiment $\mathcal{E}$, we used a total of 6 payoff structures, not all of which were symmetrical for agents interacting within them. One of the structures (situation b in particular) punished anything but cooperation, another actively encouraged defection (situation c). Yet another was completely neutral (situation d). The final two, e and f, are based on Sen's 'Assurance Game,' documented in Williams (1990). The payoff structures are given in appendix A. The experiment was run twice, once with no random cooperators, and seven each of the three dispositions; once with the same structure as experiment $\mathcal{D}$ (6 random, 5 each of the three dispositions). Each was run for 250 loops.

Summary of results

The league table for the first part of the experiment, including random agents, is shown in table 7.3. As can be seen, the mean score is higher than that for experiment $\mathcal{D}$. This is partly due to the different payoff structures, which were sometimes more generous in allocating payoffs to agents than the classical structure given above (see appendix A). In this experiment, agent S's low score can be explained by the fact that S did not interact with many other agents. Conversely, E enjoyed many interactions.

Ranking	Name	Disposition	Fitness
1	H	Pessimist	233
2	F	Random	213
3	J	Pessimist	199
4	Q	Realist	157
4	R	Realist	157
Mean			142
Minimum	S	Realist	64

Table 7.3: Summary of results: experiment $\mathcal{E}$, part one (with random agents).

Ranking	Name	Disposition	Fitness
1	H	Pessimist	372
2	R	Realist	294
3	O	Optimist	283
4	E	Realist	259
5	D	Pessimist	257
Mean			188
Minimum	F	Realist	58

Table 7.4: Summary of results: experiment $\mathcal{E}$, part two (without random agents).

Table 7.4 shows the results for the second part of the experiment, which included no random agents. The scores are higher than before. This is because agents had the chance to build stable coalition structures without random moves from those with whom they interacted. A cursory examination of the results shows that H enjoyed many interactions, and cooperated approximately 80% of the time, thus allowing partners to benefit from a greater payoff. Where H *did* defect, this made no difference, because the situation was d (see appendix A), or the situation merited it (situation c). Thus, despite agents having no *clear* method of judging the situation, knowing the payoff structure allowed them to approximate fairly sensible behaviour in these situations. This was an unforeseen result, and is quite promising.

The Introduction of Purpose

Until this stage, agents wandered around the grid at random, moving North, South, East or West with equal probability. Experiment $\mathcal{F}$ introduced the concept of *purpose* to the agents in the grid. The first part of the experiment gave agents some limited control over their movements: they could *influence* the direction they wished to move in. This actually mirrors reality quite well, since, however much we might want to do things, sometimes our choice is constrained, the same is true, but more so, for artificial agents. The method by which they influenced direction was simple: if they favoured moving in a particular direction, it was given a weighting. Normally, there is a 25% chance of moving in any particular direction. Under the weighting scheme, this is increased to 50% for the direction favoured, 10% for the opposite of that direction, and 20% each for the other two. For example, should an agent wish to move North, then the chances of her doing so are 50%. The chances of being forced to move South are only 10%, whilst the chances of being forced to move East or West are 20% each.

We performed this experiment in two parts once again. The second part contained 7 agents each of the three dispositions. From this, it was hoped that we would be able to ascertain which agents were more popular than others. The first part of the experiment introduced the random agent once again, with 6 random, 5 each of the three dispositions. From this, we hoped to see how 'popular' the random cooperator was to each of the three dispositions.

Initial configurations for each of the two tests were: agents had a memory span of 10 interactions, startup values were random, agents could 'see' two spaces only in all directions (to allow them to make their choices). The payoff structure was fixed to situation a (see appendix A). Some sample snapshots of the PlayGround at different times for experiment $\mathcal{F}$, part one, are given in appendix B. These show clearly how grouping behaviour is formed.

Summary of results

The league table of results for the first part of experiment $\mathcal{F}$ is shown in table 7.5. Once more, the scores are higher than in the previous experiments. An interesting observation concerns the number of random agents that are present in the table. It was expected that with directed behaviour, trusting agents would have even more of a chance to form stable coalitions. What appears to have happened is that the random

Ranking	Name	Disposition	Fitness
1	B	Random	302
2	L	Realist	279
3	F	Random	266
4	D	Random	262
5	K	Optimist	260
Mean			187
Minimum	C	Random	75

Table 7.5: Summary of results: experiment $\mathcal{F}$, part one (with random agents).

agents (who also *move* randomly) have taken advantage of this, moving at random into and out of groups and 'free-riding,' defecting more than cooperating, thus notching up a high fitness.

The league table for experiment $\mathcal{F}$, part two, is shown in table 7.6. Here, the expected results come through. The mean fitness is much higher than in previous examples, and the top fitness is very high indeed. What happened here was that L found itself at the centre of a very stable group, and thus interacted many times with its neighbours. It never defected, and so the neighbours chose to stay near it.

After 100 iterations, the agents were redistributed around the PlayGround by hand. Once again, however, a group quickly formed around L. This shows that high fitness values can be attained through cooperation as well as defection. This is a promising result. It also shows that group formation can help group members attain high fitnesses. In fact, all of the agents in the top five were in the first group, and after reorganisation, all but O returned to group together again.

The ability to see further

One final test was run with the same initial configuration (including random agents), but with agents able to 'see' 4 spaces, double what they could before. From this, we hoped to see the effect of more environmental awareness in trusting agents. This is labelled experiment $\mathcal{G}$.

Summary of results

Ranking	Name	Disposition	Fitness
1	L	Realist	468
2	Q	Pessimist	438
3	S	Pessimist	408
3	H	Realist	408
5	O	Pessimist	333
Mean			277
Minimum	M	Realist	90

Table 7.6: Summary of results: experiment $\mathcal{F}$, part two (without random agents).

Table 7.7 shows the league fitness table for this experiment. Whilst they are roughly comparable to those for $\mathcal{F}$, part two, they are a little lower. The reason for this is that the experiment was run over fewer iterations than that for $\mathcal{F}$. We can thus compare the results relatively, and find that being able to 'see' further did little to affect the outcome of the experiment. There may be many reasons for this. For example, being able to *see* that far does not mean that agents were able to actually *get* that far: others may have got in the way, they may have been coerced in other directions before getting where they wanted to go, and so forth. Indeed, it may have been to the agents' detriment, since they could see trusted others 4 spaces away one move, then more trusted others 4 spaces away in a different direction after another move, resulting in them continually changing direction, and getting nowhere.

Survival of the fittest

The introduction of evolutionary pressures into the system is perhaps the most difficult to justify. This is because the PD does not lend itself to giving high fitness values to those who are 'suckered.' As Behr (1981) pointed out, in order to 'do well,' i.e., score more points than your opponent, you have to defect more than they do. The result is that those strategies who take risks and try to educate non-cooperators will lose out — in that sense trust, or at least risk-taking, is not evolutionarily successful. On the other hand, once a mutually trusting relationship has been built up, interactions are rewarding to both parties, and fitness quickly increases. Thus, trusting cooperators who are able to 'stick together' should do well in evolutionary terms.

Ranking	Name	Disposition	Fitness
1	A	Optimist	426
2	N	Realist	375
3	K	Realist	337
4	L	Realist	321
5	S	Pessimist	314
Mean			229
Minimum	D	Optimist	54

Table 7.7: Summary of results: experiment $\mathcal{G}$, part one (no random agents).

Experiments $\mathcal{J}$ and $\mathcal{K}$ attempt to shed some light on these ideas. In both, an evolutionary step is taken every 25 moves (see below). Initial configurations for the experiments are as follows:

- Experiment $\mathcal{J}$:
  7 agents of each disposition (thus, 21 agents in total) on the grid, random initial configuration, movement directed, agents able to 'see' 2 spaces, payoff structure a only.

- Experiment $\mathcal{K}$:
  6 random cooperators, 5 each of the three dispositions (again, 21 in total), random initial configuration, movement directed, agents able to 'see' 2 spaces, payoff structure a only.

The evolutionary step is quite straightforward, and follows the following pattern:

```
get the average fitness of all the agents
split the agents into two camps:
   1. those whose fitness is below average
   2. those whose fitness is above average (or equal)
replace the strategy of the bottom agent in list 1 with
  the strategy of the top agent in list 2
replace the strategy of the next up in list 1 with the next down
  strategy in list 2
```

```
etc, until we reach the top of list 1 or the bottom of list 2
```

Thus, if the top agent in list 2 is a realist truster, and the bottom agent in list 1 is a pessimistic truster, then we will have two realistic trusters at the end of the run.

This version of evolution does not count for how many interactions agents had, what kind of risks they took, who they interacted with, and so on. It is, thus, fairly harsh. It still provides some 'evolutionary' measure with which to measure the progress of our agents. More realistic evolutionary techniques will be added to the testbed in the future.

Summary of results

League tables are not used here, since as each evolutionary step is taken, all fitness values are reset to zero. What can be discussed simply is the number of agents of each strategy and disposition at the start of the experiment, and the number of each at the end. With these simple results it is possible to ascertain the relative successes and failures of each strategy.

As was mentioned above, this method of evolution is perhaps a little unfair, since you are 'bred out' if you do not interact, or interact very rarely, since no interactions means a low or zero fitness. This particularly applies if movement is random, since agents have no way of influencing their final fitnesses. For this reason, it was decided that two experiments $\mathcal{H}$ and $\mathcal{I}$, which would have been identical to $\mathcal{J}$ and $\mathcal{K}$, but involving random movement, were not run, and are not recorded further here.

The results from experiment $\mathcal{J}$ are as follows. At the start, there were 7 optimist trusters, 7 pessimists, and 7 realists. Following 112 iterations (the number of iterations was limited by available memory space for one agent, U, whose memory became full), there were 9 optimists, 7 pessimists, and 5 realists. Optimists appeared to be more successful. Indeed, at the end of the run, only one of the original optimists, E, had changed strategies (to become pessimist). One realist (L) became an optimist, as did two pessimists (T and Q). These results are inconclusive, but serve to demonstrate that there is little to choose between the strategies. The story for experiment $\mathcal{K}$ is that at the start, there were 6 random agents, 5 each of optimists, pessimists and realists. At the end of the run (125 iterations), there was just one random agent (E, which had stayed random throughout), there were 8 realists, 9 optimists, and 3 pessimists. Clearly, directed movement pays off: since random agents moved randomly, they did not benefit from being able to form stable coalitions with others, and so did not gain a high fitness. Consequently, they were lost from the society.

It is perhaps too early to conclude that random behaviour is evolutionarily poor as a strategy, but an interesting parallel can be found in human society — people do not generally seek out those who behave irrationally, or at random. Consequently, such irrational people are shunned and do not interact with others as much as they could. Here again, the results from the testbed bear an interesting approximation to human society.

7.4.4 The Addition of New Strategies

In the experiments thus far, the strategies used were trusters (of three dispositions) and random cooperators. The next two experiments, $\mathcal{L}$ and $\mathcal{M}$, provided a new strategy to test trust with — that of Tit for Tat. In experiment $\mathcal{L}$, we had 6 Tit for Tat, 5 trusters of each of the three dispositions. Movement was directed (sensory limit 2), and evolution occurred every 25 iterations. In $\mathcal{M}$, the same combination of agents was used, but with random movement and no evolution.

Summary of results

The results from experiment $\mathcal{L}$ again involve no league tables. At the start, there were six Tit for Tat strategies, five optimists, five pessimists and five realists. At the end, there were two Tit for Tats, nine optimists, four pessimists and six realists. Once again, optimism seems to be a relatively successful strategy. It seems surprising at first glance that Tit for Tat does not achieve success. In fact, movement for Tit for Tat was a little confused, and did not follow the same rules as for trusting agents — the Tit for Tat agent looked around and moved towards the first agent it saw who cooperated with it last. This is more limited, and less forgiving, than the movement style for trusting agents, and because of it Tit for Tat lost out, despite being in a fairly cooperative environment (indeed, none of the Tit for Tat agents got the chance to defect, since all interactions with them were cooperative). The trusting agents were simply more able to 'forgive' (to use Axelrod's term (Axelrod, 1984)) those who may have defected last, but were still trusted, and it appears to have paid off.

Experiment $\mathcal{M}$ can be summarised in a league table, and this is given in table 7.8. In 'fairness' to agent C, who obtained the minimum fitness, it also had fewer

Ranking	Name	Disposition	Fitness
1	Q	Realist	200
2	A	Tit for Tat	129
2	D	Tit for Tat	129
2	N	Pessimist	129
5	B	Tit for Tat	126
Mean			112
Minimum	C	Tit for Tat	63

Table 7.8: Summary of results: experiment $\mathcal{M}$.

interactions than other agents. Here, movement is random, and the scores are consequently lower. In addition, Tit for Tat shows its success as a strategy in the Prisoners' Dilemma. In fact, the scores were all very close to each other, hovering around the mean point. Agent Q interacted with a relatively large number of other agents, but defected every time. Again, with a society of communicators, its success would be limited. As it is, it does well. It is safe to conclude here that random movement presents itself as something of a problem for the agents in this testbed, and that directed movement allows the formation of stable coalitions which enable agents to cooperate without fear of being 'taken for a ride.' For trusting agents, this is important, particularly in the first stages of a relationship, where little is known of the other, and the agents can in principle be taken advantage of. Communication amongst members of society should prevent even that. A future implementation will include such communication.

7.4.5 Adjustment of trust

The strategy used for the adjustment of the trust values after an interaction is described above, and follows the rules:

1. **Both cooperate.**
   Basic Trust increased by 1%
   General Trust in other increased by 10%

2. **A cooperates, B defects.**
   Basic Trust decreased by 1%
   General Trust in other decreased by 10%

3. **A defects, B cooperates.**
   Basic Trust increased by 5%
   General Trust in other increased by 1%

4. **Both defect.**
   Basic Trust decreased by 5%
   General Trust in other decreased by 10%

These rules, however, pose their own problems: as the trust value to be adjusted tends to zero, the adjustment decreases, and the trust value never:

1. Increases above zero if it starts out below it.

2. Decreases below zero if it starts out above it.

This is a considerable problem for the formalism. The experiments use this particular strategy because it provides ease of implementation and for the most part agrees with the discussion presented in chapter 4, also, the problem was not evident until the implementation was run with the previous experiments — another indication of the use of such an implementation. Another strategy is to increase or decrease by some fixed value, say 0.01 for basic trust, and 0.05 for general trust. Using this simple strategy, we performed one more experiment, $\mathcal{N}$. This was identical to the second part of $\mathcal{F}$, with 7 agents each of optimists, pessimists, and realists, and with directed movement and a sensory limit of 2 spaces, and used the new strategy for the alteration of trust.

Summary of results

The league table of results is shown in table 7.9. Not surprisingly, considering movement is directed, the scores are very high. There were some extremely low scores, all of them amongst the random agents. This is again due to their lack of directed movement and subsequent lack of ability to build stable coalitions. In this experiment, K behaved as L in experiment $\mathcal{F}$ part two, and was at the centre of a very successful group. It therefore never needed to defect to build a large fitness. The other four

Ranking	Name	Disposition	Fitness
1	K	Optimist	519
2	S	Realist	459
3	H	Optimist	399
3	J	Optimist	399
3	N	Pessimist	399
Mean			262
Minimum	D	Random	23

Table 7.9: Summary of results: experiment $\mathcal{N}$.

leaders, S, H, J and N were also part of that group. This is more evidence of the efficacy of groups, and the ability of trust to assist in their construction.

7.5 Discussion

7.5.1 The PlayGround

The PlayGround provides an ideal testbed for the implementation of simple trusting agents. It provides user-control of agents positions, the workings of evolution, the possible payoffs agents can receive, and the addition of extra strategies, trusting or not, into the society. In addition, it provides the important concepts of physical distance and proximity to other agents, a knowledge of who other agents are (if they have been interacted with) and that of a society of agents. All of these have been lacking in previous Prisoners' Dilemma experiments, and all of them are important to trust.

The results obtained from the experiments are indicative of the fact that the formalism does indeed mirror quite closely the behaviour of trust when it is embedded in social agents. In itself, this conclusion is beneficial to further work on the formalism, since it shows that such work would be practicable.

The experiments provided pointers to some of the problems that remain with the formalism, particularly as regards the alteration of the trust value following interactions with other agents. This is an area of critical importance to the future acceptability of the formalism and the concept of artificial trust since within this alteration lies much of the inherent adaptive ability of trusting agents. Further work will be necessary to arrive at a suitable strategy, or set of strategies, for the alteration of the trust value. Initial results with the experiments show that the strategies used for the alteration of trust do result in sensible behaviour.

7.5.2 The Findings

Despite some limitations, the findings of the experiments are relevant and provide insights into the usefulness of trust as a decision making strategy. Some interesting and detailed results have come from the PlayGround. There are reasons for this. Firstly, the concept of a society, with the freedoms that entails, was embedded in the experiment (albeit in a limited fashion), and this additional complexity served to provide extra information and capabilities to agents. This is reflected in the results obtained, especially when movement was directed. Secondly, the implementation itself was more complex, and contained implementations based on the more up-to-date version of the formalism, than the tournament experiment.

In the tournament, although the number of strategies involved was considerably higher than for the society, the obtained results were limited in their applicability — all that was shown was how trust fared *against* other strategies. The societal implementation attempted to go beyond that, taking into account the limitations of the PD as a simulation technique, and provide a notion of interactions *with* other agents. This was aided by the inclusion of several different payoff structures for situations. The results from this are thus more pertinent. Less applicable at present are the results obtained from the evolutionary test, because of the limitations discussed above.

7.5.3 Society

The concept of society is crucial to an implementation of trust. If society were not present, trust would have limited application (see chapter 1, also chapter 3). Indeed, it has been argued that society needs trust just as much as the converse. Accepting this as the case, it is exciting to note that, when directed movement was enabled for agents in the PlayGround, 'mini-societies' grew up, in the form of groups of agents migrating towards those they trusted, and within these groups, the amount of trust increasing and staying high because of repeated interactions with trusted others. Here we can see primitive groups and societies forming. This is an important avenue for further work.

7.5.4 Observations

There are several observations that can be made following these experiments and the subsequent discussions. Among them are the following:

- The shorter the memory span, the more alike dispositions become, although trusting behaviour still occurs (trust is still altered).
- A memory span of 1 means all dispositions are the same.
- A memory span of 0 means the agent is random.
- Optimists are quicker to cooperate than Realists, following repeated risk-taking by the other agent. In the experiments above, they appear to be more successful for that.
- If their initial trust is too low, Pessimists will never cooperate.
- Trusting behaviour can be educated. This is analogous to Axelrod's finding that cooperation can be encouraged, particularly by Tit for Tat (Axelrod, 1984).
- Some of the trusting agents in the experiments behaved in a fashion remarkably similar to that of human agents. The primary examples concern what Deutsch calls pathological trust — trust where there is clearly a great risk and cost involved. An example of this is shown in experiment $\mathcal{A}$ on page 148.
- In less extreme cases, artificial trusting agents behave in ways similar to those of human trusters, cooperating when trust is high and risks low, and *vice versa.*
- Where directed movement is involved, trusting agents appear to be able to build stable coalitions. These coalitions are beneficial to their members, since cooperation is almost assured within them.

Although some of these observations are of limited application, they are nevertheless indicative of the claim of this work that trust can be formalised, and that the formalism can be embedded in artificial agents to give trust-like behaviour.

7.6 Summary

The primary aim of this work was to formalise and operationalise a concept of trust. The formalism was at once relatively straightforward to work with and easy to implement. Implementations of the form outlined above are clearly possible, and trust appears to give agents an extra decision-making property. It is too early to say whether this property is to the good of trusting agents. Clearly, more detailed experimentation and implementations are necessary to ascertain this. The following chapter discusses some of these. A next step is to involve the human being[4] to a much greater extent, possibly in the form of a player in a strategy-type game involving negotiation and risk. With such a platform, dispositions and various decision-making procedures can be tested in detail.

The experiments and implementations that have been carried out serve two major purposes. Firstly, they provide confirmation that the formalism can indeed be embedded within artificial agents, and that the agents can reason using this limited form of trust. Not only that, the subsequent behaviour of these trusting agents is remarkably similar to that of humans, particularly when the more pathological forms of trusting are observed (such as when a very high truster continually loses out when interacting with an unconditional defector or a low truster). These insights are very promising. Secondly, they provide pointers to further work, both in terms of additions and enhancements to the formalism, and for future experiments and areas of implementation. The following chapter discusses some of these avenues.

[4]There are in fact at least two ways of doing this. The first invloves incorporating trusting agents into some kind of game that humans can play, and is suggested here. Another is to involve the agents in more serious implementations, for example as 'personal assistants' in some CSCW application.

Name	Strategies Used	Situations Used	Movement Style	Evolution Present?	Number of Iterations
$\mathcal{D}$	6 Random 5 O/P/R	a only	Random	No	250
$\mathcal{E}$	No Random 7 O/P/R	a, b, c, d, e, f	Random	No	250
$\mathcal{F}$ (part 1)	No Random 7 O/P/R	a	Directed (2)	No	250
$\mathcal{F}$ (part 2)	6 Random 5 O/P/R	a	Directed (2)	No	250
$\mathcal{G}$	6 Random 5 O/P/R	a	Directed (4)	No	250
$\mathcal{J}$	No Random 7 O/P/R	a	Directed (2)	Yes (every 25)	112
$\mathcal{K}$	6 Random 5 O/P/R	a	Directed (2)	Yes (every 25)	125
$\mathcal{L}$	6 TFT 5 O/P/R	a	Directed (2)	No (every 25)	125
$\mathcal{M}$	6 TFT 5 O/P/R	a	Random	No	200
$\mathcal{N}$	6 Random 5 O/P/R	a	Directed (2)	No	128

Table 7.10: Summary of the types of experiments carried out using the PlayGround

Chapter 8

Future Directions

8.1 Discussion

The work in this thesis introduces:

1. A formalisation of the concept of trust which allows precise discussion and experimentation to be performed.

2. An implementation of trust, using the formalism, in an embodied agent.

The formalism not only presents opportunities for such an implementation, but also provides areas such as social psychology with an precise means of discussing trust. The work is, however, introductory, and a major goal is to spur further research in the area. As has been mentioned elsewhere, both in this thesis and other work (Marsh, 1992, 1993, 1994a, 1994b), there are some caveats to be borne in mind.

- The formalisation presented here is not the only way to formalise trust. It is one way, and it is presented for discussion.

- Trust is not a stand alone option. The examples and implementations presented here use trust without other considerations because it is trust that we are trying to simulate. The decisions we make in everyday life, however, take into account far more than trust (Danielson, 1992a; Agre & Chapman, 1987; Chapman & Agre, 1987; Luhmann, 1979; Kiss & Reichgelt, 1991). Not least, ethics, morals, cultural pressures, and the wants and needs of the decision maker are of importance. The argument that trust plays a major part in such decision making does not detract from these other considerations.

The presentation of a formalism for trust can be seen as one step along the road to an embodied agent which deliberates in a sensible, moral, ethical manner whilst taking such important factors as trust into account. Work in areas such as Artificial Morality (Danielson, 1990; Danielson, 1992b; Danielson, 1992a) and the semantics of desires (Kiss & Reichgelt, 1991) can be seen as steps in the same direction.

This chapter provides detailed overviews of where future work may be feasible with regard to trust in embodied artificial agents, and where the formalism can be extended to reflect better the actual workings of trust. We go further since, as was discussed above, trust cannot really be taken as an individual concept in the world, neither do we propose that the formalism described in this thesis is the final word on the subject. Rather, it is one of the aims of the work to promote future research in the area, leading to better formalisms, or different views of trust, and to different types of representation. To that end, several of the sections in this chapter provide alternative ways of thinking about trust, and developments of the formalism provided here which are of use in areas such as social psychology, DAI, and CSCW. In addition, the concept of legal systems to augment the decision making process is discussed (section 8.5). This takes into account some of the problems of the formalism.

8.1.1 Overview of the Chapter

The chapter is composed of several sections, each of which applies to a separate consideration of the formalism or its implementation. Each section attempts to discuss where the use of trust, and in particular the formalism, can take us from where we are, presenting a series of questions which will for each topic suggest avenues of future research. The major area of consideration is DAI, thus the first section here is concerned with applications of trust in DAI, in addition to giving agents capabilities of reasoning using trust. Section 8.3 examines other areas in computing, those of CSCW and Human Computer Interaction, within which trust could be of use. As mentioned above, there are other means of considering trust, other formalisms which could be applied, and so forth. Section 8.4 presents a view of Japanese society — a society based on trust which actually exists. It is based on Yamamoto's 1990 paper, and provides an interestingly different view of trust, and a society built around it. Legal considerations and extensions are given in the following section, based primarily on Minsky's work with law-governed societies (Minsky, 1991b; Minsky, 1991a). The final two sections consider alternative formalisms, extensions to the present formalism, and other aspects which this research has uncovered.

8.2 Distributed Artificial Intelligence

It is no surprise that there remains much to be done with regard to DAI and trust. This thesis presents a basic formalism and some experimentation to justify the use of trust, particularly in cooperative situations. DAI, however, goes further than that. As a subject of study, it provides a wide vista of possibilities, from artificial reasoning (Georgeff, 1984; Chapman & Agre, 1987; Hayes-Roth *et al.*, 1991) to a model of complex social phenomena (Castelfranchi *et al.*, 1991; Numaoka, 1991; Drogoul & Ferber, 1992). The trust discussed so far has to some extent been concerned with an extension to the reasoning capabilities of agents. This section discusses other areas where a knowledge of trust would provide extra capabilities.

8.2.1 Modelling Human Trust

One aspect of DAI is the capability of societies of intelligent agents to model in some way more 'human' organisations, thus allowing us to advance understanding in such fields as sociology and organisational theory (Bond & Gasser, 1988). What has been presented so far in this work is basically an extension to the reasoning capabilities of artificial agents. In addition, we have mentioned more than once the ability of the embedded formalism to model human trust closely. At the present time, the formalism is relatively simple in this respect. This is a strength since, as the formalism works and can be implemented, its simplicity allows detailed examination of the concept of trust. The preliminary results obtained from implemented agents suggest that their behaviour in terms of trust is, in fact, as might be expected (see chapter 7). However, for the formalism to provide any serious answers to questions concerning *human* trust, much remains to be done. Below, we present an outline of some research that can be carried out in the near future.

Questions of Trust

The questions presented here are not based on any one piece of work; rather, they arise, implicitly or explicitly, in much of the literature concerning trust. The fact is, trust has had a meagre following as a mainstream research topic, and most of the work that has been done has been of a relatively informal manner. Thus, some aspects have been blurred in the search for an overall viewpoint of trust (for an excellent discussion of this, see Luhmann (1979), which is followed up in Luhmann (1990)). One major question that has attracted consideration is what trust actually *is*. There are several definitions (see chapters 1 and 3). The fact that there are several should be evidence enough that the question has not been answered satisfactorily to now. Deutsch's 1962 definition appears to be the most acceptable at present. Indeed, much of the formalism presented here is derived from Deutsch's work, particularly the first two parts of his definition (see page 32). We disagree with the third part, since we argue that, in fact, Va^- may be equal to Va^+ and still need a trusting choice. Were Va^+ greater than Va^- the choice of the ambiguous path would not, argues Deutsch, be trusting, as the individual would not be taking a trusting decision, rather, he would be taking a risk, or gambling: "one trusts when one has much to lose and little to gain" (Deutsch, 1962, page 305), whereas "one gambles when one has much to gain and little to lose" (*ibid.*, page 305). We disagree only in that, when gambling, one generally knows the odds against winning. When trusting someone, one puts oneself in their hands, as it were — the final result is up to them. Trust involves risk, uncertainty, and the actions of another. Unless we count the hand of fate, gambling involves only the first two of those. Thus, whether or not the positive outcome is greater than the negative, when we place our fate partly or wholly in the hands of others, we make a trusting choice.

This only serves to highlight the disagreements that can arise in a discussion of trust. By observing different behaviours for agents under different circumstances, with slightly differing definitions in mind and adjustments to the formalism to take these into account, we can begin to reach a definitive answer to our first question:

1. **What is Trust?**

   It is useful to consider trust removed from other considerations — ethics, emotion, and so forth. This allows us to see the results of trusting behaviour alone, without complicating matters. It is, however, a simplifying assumption. Human beings at least make trusting decisions in the light of many other considerations, morals, ethics, and emotions. The question becomes, how much do we lose by isolating trust? Is trust separable from humanity as an independent entity, or does it depend on other aspects of our selves? In many ways, this question is similar to that posed by Descartes, the Mind-Body problem. While answers to this are predominantly philosophical, DAI and AI can and do advance scientific knowledge to the extent that we may be able to answer the question empirically. The same goes for the next question:

2. **Is Trust independent of other [human] considerations such as morality, emotion, and ethics?**

   Which in turn leads to the asking of another:

3. **Are human traits such as morality, ethics, and emotion independent of Trust?**

   With careful experimentation, answers to these questions can be provided. It is not difficult to see the direction experiments would take. The work presented here is a first step in that direction, and a practical proof that an implementation can be carried out.

   We present some less challenging questions here:

4. **Does a linear formalism do justice to trust? Would more complex representations capture the spirit of the phenomenon more precisely?**

5. **Cooperation aside, where else does trust play a part? How can we represent such in the formalism?**

6. **Is Trust a necessary part of human behaviour? What if trust was absent? (i.e., what if complete distrust was all there was?)**

All of these are interesting questions in their own right. They do, however, present the beginnings of a detailed research strategy which can build upon the work presented in this thesis and develop a detailed theory and practice of trust which is of use in the social sciences. In this, DAI can play a major role. Naturally, there are other questions; some are known, some are intuitive, some will only appear as the research proceeds.

A Research Strategy

Clearly, the questions discussed above provide a direction for future research. The formalism introduced in this work was intended to clarify the situation with respect to the many views of trust, with applications in embodied agents. Both objectives have been demonstrably fulfilled. However, the limitations are clear — trust alone is not a suitable decision tool for an agent, rather it augments the agent's repertoire of decision making capabilities. The same is true for human agents. If we were to use the ideas given here to provide insights into human trust, the formalism would have to be altered a great deal. At present, it is weak in several respects. Not least, it is guilty of glossing over certain key aspects of the makeup of trust, such as risk, competence and cost-benefit analysis. This is justifiable since we were attempting a first step to a formalisation of trust. This being the case, something which behaved like trust was necessary before we could begin to improve upon it. Improvements have been carried out, resulting in the formalism presented in this thesis (which is considerably different from that first proposed in Marsh (1992), and Marsh and Thimbleby (1992)). Acknowledging this, more needs to be done. A deeper understanding of the underlying aspects of trust, especially competence and risk, is necessary. In addition, the formula for situational trust is weaker than we would like: there are too many exceptions to the rule. These need to be ironed out before we can present the formalism as 'how trust *really* works.' The formalism has, in some respects, grown to suggest itself in that light. It promises much in that respect, without over-generalisation.

If the formalism can be altered and refined in this manner, then some of the questions above will have been answered, not least question 4. In addition, other questions will have been addressed. Questions such as 'is *any* formalism good enough to represent trust?' In other words, is trust a phenomenon which can be represented in a formalism, or is there too much to it? Assuming that trust can be formalised, we can proceed to trying to address the deeper issues discussed above. If trust cannot be formalised, then the work presented in this thesis is a step towards greater understanding of what cannot be done with trust, and perhaps other social human-centered phenomena, such as the emotions.

8.3 Computer Supported Cooperative Work and Beyond

The definition of agent used in this thesis does not preclude humans, thus, away from 'pure DAI,' the field of CSCW, treating humans and the computers they use as agents, presents itself as an ideal application within which trust will be of use. Indeed, work in progress (Thimbleby *et al.*, 1994; Jones *et al.*, 1994) suggests that trust is a practicable and novel addition to the understanding we have of cooperative work involving computers. This work should provide worthwhile predictive results in this area.

In any case, this work shows that trust as we see it here is not confined to DAI, or even to fields such as sociology or social psychology — it has applications in any areas which are inherently 'social' (a concept which is becoming more accepted as time goes on — cf. Communications of the ACM, January 1994). The field of Human-Computer Interaction (HCI) is another example of this — interactions proceed between human and computer (much as in CSCW, without the wider social aspect), and the aim is to make the computer as acceptable as possible to the user. This could simply be a matter of setting up a windows workstation such that the windows are in their correct place for that user. Much of the work in HCI can be seen as trying to increase the amount of trust between user and machine (Muir, 1987)[1].

The exclusion of humans as agents does not lessen the scope of trust a great deal. There are areas where *computers* and computers work together in a social fashion. For example, the telephone network in certain industrialised countries is almost entirely computer-controlled. Computers decide, according to various routing algorithms, which way to send messages around the network, and so forth. This can lead to unexpected, *emergent*, behaviour (e.g., the AT&T long distance collapse in 1990 (Neumann, 1992)). That the computers work together socially is perhaps a contentious and nonsensical argument, but emergence among socially complex organisations is undoubtedly a powerful factor in their overall behaviour (Chapman & Agre, 1987; Wavish, 1991). When we accept that, we can accept that trust plays a part in such social interactions — it is plausible, then, that routing algorithms can be designed using trust in other nodes (local or more distant) in the eyes of a particular node — a low trust means that the message will probably be sent in another direction, or in short bursts, to minimise loss should errors occur. Clearly, this is an interesting avenue of work to undertake.

There are other examples of areas where an inherent social nature allows the application of trust in such a way that behaviour may improve or efficiency increase.

[1] I am grateful to Harold Thimbleby for pointing this out.

Indeed, the area of Human Computer Interaction has in the past received some input regarding trust from other areas (Muir, 1987; Lee & Moray, 1994). The identification of other potentially fruitful areas is an interminable task, and one which is unnecessary for this work. It is undoubtedly the case that such areas do exist, and further work could potentially discover them.

8.4 *Wa* — A Moral Society

Yamamoto (1990) presents a discussion of harmony, or *Wa*, in Japanese society — a society based on mutual basic trust (Yamamoto, 1990). This section takes the basic ideas of relationship-contexts from that paper, and introduces the basics of a formalism that may be based on them. Yamamoto suggests that in Japanese society there are three contexts of relationships, each of which depends on the amount of mutual basic trust that can be expected of the participants. They are:

> *Context 1:* The presumption of (general) mutual basic trust is beyond reasonable doubt. The relations between members of close-knit families, between intimate friends, and between lovers are examples.
>
> *Context 2:* The presumption of (general) mutual basic trust is reasonable, but it is not beyond reasonable doubt. Examples are relationships between neighbours and casual acquaintances.
>
> *Context 3:* The presumption of (general) mutual basic trust is unreasonable. A meeting with a person on the street for the first time is an example; the first meeting between a Japanese and a foreigner is, for the former, a paradigm.
>
> Yamamoto, 1990, page 463.

At any time, a person may be in a context 1 relationship with some, and 2 or 3 with others. It is also possible that "a Japanese may regard her relationship with a single individual as being context 1 *in some respects*, but context 2 or 3 in others." (Yamamoto, 1990, page 463). The simple formalism we present here takes these contexts and allows agents to reason using the context they believe themselves to be in with regard to other agents. Thus, in a context 1 relationship, cooperation is assured (or should be). In context 2, more deliberation is needed, and in context 3, some form of guarantee (a contract enforceable by law, for example, or the safety net) is needed.

Clearly, the concept of *Wa*, or harmony, may be of use in extending our notion of trust, not least because it allows trusting agents to develop behaviour appropriate to the conditions in which they find themselves without searching for that behaviour (a process which may be taxing for more simple agents, both in terms of time and processing power). In other words, if an agent perceives itself to be in a particular relational context with another, in some instances trust need not be considered — it is either not accepted in that context or it is to be assumed as 'high.' Only in certain contexts (notably context 2) need the agent consider trust at all seriously. This does not detract from the idea that trust is present, it is simply a form of filter which allows an agent to consider trust in an implicit way, whilst acknowledging its presence, or the lack of it. Future research in such a direction may help answer some particular questions, for example to what extent trust needs to be considered in a particular situation, and which extent in which particular situation is the most efficient in terms of speed and processing power or space. It is worth mentioning that, simple as the formalism is, there are likely to be simpler agents, or agents with little time or space to consider it. The question then becomes, how little of trust can these agents get away with knowing and still benefit. *Wa* suggests ways around that particular dilemma.

There is something else, however, which the concept of *Wa* suggests. In particular societies, trust perhaps has different meanings. The formalism developed in this work is inherently based on Western cultural ideas and ideals. Without digging too deep into the philosophical aspects of this, it seems clear that other cultures or religions may disagree with the idea. This being the case, it is clear that alternatives are possible. *Wa* is one of those alternatives in that sense — it provides a different way of looking at how trust works, suggesting new lines of thought and avenues of research. This is no bad thing, as it serves to reinforce the argument that the formalism presented here is not 'how things must be.' In addition, it gives weight to the formalism because it shows how we can combine the two to provide a stronger consideration of trust.

The idea of different formalisms and ideas upon which to base the concept of trust is discussed further below, in section 8.6, which also suggests different value bases for the present formalism. Before that, we proceed with a deeper discussion of the concept of legality in systems which use (or don't use) trust as a basis for decisions — this being a concept which has been mentioned several times during the thesis (see in particular section 5.3).

8.5 Law-Governed Societies

The previous section mentioned that it may be possible to reinforce the formalism presented in this thesis by augmenting it with other concepts to allow deeper consideration, or more shallow consideration, of some aspects of trust without losing any expressive power (and perhaps gaining some). This section is inspired by Minsky's work on Law-Governed Systems (Minsky, 1991a; Minsky, 1991b). This work is based in work in distributed systems, and provides a Law-Governed Architecture within which distributed systems operate. We provide an overview of Minsky's work here, and proceed to show how it can be allied with trust to provide an extremely robust decision-making agent in a complex distributed environment. We go on to suggest several applications within which to test the ideas presented.

Minsky's notion of law-governed systems (Minsky, 1991b) is concerned with the management of large scale evolving systems in software engineering. The law of such a system is "an *explicit* and *strictly enforced* set of rules about the operation of the system, about its evolution and about the evolution of the law itself." (*ibid.*, page 285). Minsky's argument is that large 'lawless' systems are incomprehensible and incomplete (unpredictable). The specification of a set of rules about "who, and under what conditions, can do what to which part of the system" (page 285) will help to rectify that situation, making the system more predictable, and more comprehensible, both in terms of behaviour and end results. In addition, because the law is strictly enforced, programmers can rely on such constraints when developing it, throughout its life-cycle. Finally, it is emphasised that the law is not a general law of software, rather specific systems should have their *own* laws, designed specifically for them, and allowed to change with them. Thus particular circumstances relevant to a specific system can be taken into account in the law. There are, however, some requirements which a law should uphold:

- A law "should have jurisdiction over both the operation and the evolution of the system it governs" (Minsky, 1991b, page 286).

- It should be explicit.
- It should be strictly enforced.
- It should be "mutable in a self-regulatory manner" (*ibid.*). Thus changes in the law should be subject to the law itself.

Minsky presents an abstract model of law-governed systems, the detailed implications of which are not relevant here, apart from the definition of the law-governed system as a triple:

(system, law, enforcer)

Where the system is a set of objects which communicate by means of messages, the law is a set of rules about the exchange of messages, and the enforcer is a mechanism that enforces the law. Leaving the field of software engineering for the moment, it is clear that such a system can be applied to DAI, where the system is the society of interacting agents, the law is a set of rules about the structure of these interactions, and the enforcer is something which enforces that law. Such a mechanism could be the society itself, using social sanctions to ensure legal behaviour (see chapter 5. See also Rosenschein (1985)). In terms of DAI, however, this is a weak answer to the problem, since the society of agents can be lax in such a role, due to its distributed nature — social sanctions can work well in tight knit societies, but more distribution presents problems in getting them to work effectively. In addition, society can be broken down (Lagenspetz, 1992). There is then a need for an explicit enforcer, which would itself be an agent, which monitors and enforces the law, and changes in it. This enforcer would be as subject to the law as any other agent, and need not be omnipotent — the fact that the law exists is often enough to prevent crimes from occurring. If the enforcer took the role of an ombudsman (with specific powers) to whom aggrieved agents could turn, this may allow trusting agents (and non-trusting agents) to perform their tasks in the secure knowledge that the law is available for use, the enforcer is effective and can be turned to when needed, and all agents know this (If it were not common knowledge that the law existed, it would lose much of its power).

It is possible to envisage two types of situation where the enforcer would need to be consulted:

1. When some contract is being drawn up. Here, agents are trying to avoid possible deadlock with recourse to a safety-net which involves the law. Contracts are drawn up which are legally binding, with compensation should they be reneged on, and cooperation can proceed. This is a passive aspect of the enforcer, but perhaps the most important (Lagenspetz, 1992).

2. A more active aspect of the enforcer is seen in the second type of situation, where aggrieved agents resort to the law to help 'punish' those who have behaved badly. In this circumstance, the enforcer takes the role of judge and jury, doling out punishments as it sees fit — for example, an agent could be declared ostracised, and further interactions with it could be made impossible, or for a lesser offence, the agent (or its owner) could be fined or excluded from the community for a specific time. Here, the enforcer takes an active role in seeing to it that the punishment is carried through.

It can be seen that the concept of the enforcer presented here is slightly different to that present in Minsky's work. Minsky sees the enforcer as something which *strictly* enforces laws, making sure nothing untoward happens. This is clearly impossible in a large society of distributed intelligent agents. In such a system, the enforcer can only act as a back-stop for agents to rely on in the event of a grievance or the need for a safety net. It is conceivable that the implementation of such a system would be fairly straightforward. The concept of an omnipotent enforcer is not, in fact, unlike the situation that occurs with contract nets (Smith, 1980), where agents bid for contracts from other agents. The agent giving out contracts can perform as the enforcer *for that particular contract.* There is still, however, an element of risk that those who did the bidding will not do the work, and it is not clear how these transgressions could be 'punished.'

8.6 Other Aspects

The formalism presented in this thesis is a starting point for work concerned with trust, particularly in Multi-Agent Systems. It is argued that, an agent who can reason with and about trust will be more capable of making reliable, informed decisions than one who simply ignores the phenomenon, or assumes it is present all the time, or, worse, who assumes it is not at all present. Since it is a starting point, however, it leaves much to be done. The previous sections have suggested further work in such areas as sociology, DAI, HCI and CSCW, in addition to mentioning certain specific augmentations of the theory of trust we have built up. In this section, we present further enhancements to or changes to the formalism as it stands, discussing various problems which are evident, and suggesting ways to address them.

8.6.1 Why Use Values?

Chapter 2 discussed the use of values in this formalism. It was argued that, whilst explicit values placed on trust pose certain problems (e.g., whose subjective, or agent-centered, value we are observing, and whether or not two agents' values that are the same mean the same things in terms of trust) the use of values does carry certain benefits. Not least of these benefits is that we are able to provide a formalism which is linear and is simple enough to understand intuitively whilst preserving the power of both the formalism and the concept of trust. In addition, we are able to experiment with different value bases, to allow us to determine which base is most representative. We are not alone in the use of values — Gambetta (1990a) has suggested that trust takes a value in the interval $[0, 1]$ (see chapter 3). That said, as far as we can ascertain we are the first to provide a formalism for use with trust, and the first to suggest an implementation of that formalism.

The use of a different value base provides some interesting questions. If trust were suggested to take a value in the interval suggested by Gambetta, no real sensitivity would be lost, since we use real numbers. So, anything in the range $[0, 0.5)$ maps to our value of $[-1, 0)$, and any value in the range $[0.5, 1)$ would take a value in the range $[0, 1)$ in our formalism. The formalism as it stands would probably have to change somewhat. One of its problems is inherent in the use of fractional numbers multiplied by other fractional numbers. Inevitably, the result is smaller than the multipliers. Whilst this works well when we consider that we have negative values to represent degrees of distrust, thus anything positive, however small, represents some degree of trust. It does not work at all when we are restricted to positive values for all degrees of trust and distrust. There are two major problems with the more restricted range of values:

1. It is less clear that we are talking about trust or distrust at any one time.

2. The formalism needs to be considerably more sophisticated to take trust and distrust into account.

The first point is to a large extent academic, since the artificial agents using trust would not notice the difference. The second point is more problematic. One solution is to convert the more restricted range to its mapped value in the larger range, perform the calculation, and map the result back. This is clearly impractical. A modification to the formalism is required. This is beyond the scope of the current work, and would need to be devised over a long period of time and experimentation. It is unfortunate that the formalism cannot be generalised across different value ranges. This problem can be addressed in the future, should the present range of possible values for degrees of trust become inadequate. This is unlikely, since any such range is infinite. In practice, however, a more large-grained range may be more useful to artificial agents. Since it remains to be seen if a complete formalism of the workings of human trust can be found, this may be largely an academic debate.

A final aspect of the use of values concerns their sign. Clearly, multiplying negative values results in positive values. This is an unexpected bonus with regard to the formalism, particularly with the formula for situational trust:

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

With a negative utility estimated, it follows that an agent would not wish to carry through this situation. If a negatively trusted agent was present, the final trust value would be positive (assuming importance is positive). Thus the *dis*trusted agent is *trusted* positively *not to* carry out whatever task is needed in this situation. This is an interesting anomaly, but on consideration seems useful.[2]

This interpretation of trust is clearly a Machievellian approach to trust. Naturally there are others. These other approaches, for example masochistic, are dependent on the algebra chosen to represent the formalism. Different algebras may well result in different approaches. See section 9.6.

Deceit

Despite all of the checks and balances supplied by the law, safety nets, and experience with other agents, there will still be the possibility that agents can and will be deceitful.

[2] Many thanks to Harold Thimbleby for pointing this out.

Here we present a brief discussion which raises some interesting questions about trust and deceit.

In practice, if the workings of the formalism are common knowledge, then any agent can make use of this to manipulate the truster into a position where they can be taken advantage of by behaving in a manner which actually *increases* trust in the short run, then reversing this behaviour in order to gain the maximum benefit for themselves at the expense of the truster. Interestingly enough, this can be seen in trust in humans, hence the success of confidence tricksters.

How can we reinforce trusting agents against such events? The intuitive answer is that this is not possible — if trust and its workings are common knowledge, then deceitful behaviour is possible. The workings of the formalism aside, there may be a method in the algorithm we use for the alteration of trust following specific behaviour. It is generally accepted (see earlier) that trust increases following positive behaviour, and decreases following negative behaviour. The amount of the increase or decrease is what is of issue here. As suggested by Rempel *et al* (1985, 1986), a sudden defection from a trusted friend can result in a drastic reduction of trust, to the extent that a lot of work is necessary to build that trust up again. This suggests that confidence tricksters can only get away with their tricks once with any one person. This does not help society as a whole, however, until the information about that trickster is disseminated. In humans, emotions such as shame or embarrassment may hinder such dissemination. We suggest that in artificial agents, this would not be the case, and that, coupled with the use of the enforcer, sanctions could be taken immediately the trick came to light. Thus, at the cost of one agent, society is protected against such strategies. Note that, as trust, such an answer can only exist in societies of agents, where communication is available between agents.

To take deceit into account, the formalism may not need to be extended, but more thought has to be put into the alteration of the trust values used by agents. One of the limitations of what has been presented in this work is that of the alteration of trust. Indeed, this is a large topic deserving much further work. The testbed developed for this work provides the ideal ground for experimentation in this area.

8.7 Groups as Entities

At several points throughout the thesis, we have mentioned the idea of a group being considered as an individual entity (see especially chapter 1). The idea of a number of 'agents' coming together to form a whole is not, in itself, an unpopular one. Minsky's Society of Mind (Minsky, 1986) is a seminal work in this particular area. There are others, relating to different aspects of groups, but nonetheless of importance. We have ignored many of the aspects of groups in this work. The reasoning is simple — we have been trying to introduce a concept which pertains to the individual in that it is the individual who makes the trusting decisions. Considering a group as an individual does little to change that — the group will still make its choice as an entity, not a collection of them, as far as that analysis goes. Since we were primarily concerned with the formalisation of trust, we were little concerned with the intricacies of group behaviour.

Since cooperation between individuals may eventually lead to the formation of groups (Argyle, 1991), trust has a large part to play in their formation (Golembiewski & McConkie, 1975). Investigation into the part of trust in forming and continuing the formation of groups is, for the reasons discussed briefly in the preceding paragraph, outwith the scope of this introductory work. Needless to say, it is a fascinating area, and one which deserves thought. It is also a complex area in social psychology, sociology and philosophy. The introduction of the trust formalism to help analyse and simulate groups in existence and those in formation and breakdown may help clarify some problems. In addition to the tools already available for analysing multi-person interaction (e.g., Markov Chains (Suppes & Atkinson, 1960), and for larger societies, Metapopulation Dynamics (Hanski & Gilpin, 1991; Gilpin & Hanski, 1991)), the trust formalism provides an aspect of *reason*. It can help provide answers as to *why* these things happen in terms of individual group members. We thus observe groups from a micro, not a macro level. This is in common with much of the work done regarding cooperation, groups, and individuals (Argyle, 1991).

8.8 Additional Future Work

During the course of this thesis, there have been many places where future work has been mentioned. This chapter has discussed much of that work. There are omissions,

Page No.	Description
84	Fix problem of zero importance.
106	Strategies for altering trust.
107	Trust in DAI agents.
138	Agent-subjective estimates in complex situations.
150	Estimations of utility.
159	Fixing evolution problem in PlayGround.
161	Societies of communicating agents.
165	Formation of societies of agents.

Table 8.1: Other work for the future

however. In the text, there have been places where we discussed problems with, for example, the current formalism, which would have to be addressed in the future. Rather than repeat these discussions, table 8.1 provides a page reference to each aspect.

8.9 Summary

The formalisation of trust presented in this work is not complete. It is a piece which is in flux, and this is as it should be (Popper, 1967; Popper, 1969). In addition to being able to work further on the formalism itself, we have identified several areas where a knowledge of, or use of trust would be practicable. Amongst these are CSCW and DAI.

The identification of these extra avenues for work illustrate the strength of the formalisation, but also the all-pervasive nature of trust. It is, indeed, present in all systems which have a social aspect. Whether or not an explicit understanding of it is necessary depends on the system, but in order to determine that, the system has to be addressed.

The next chapter draws some general conclusions and summarises the work that has been presented here.

Chapter 9

Conclusions

9.1 Introduction

The thesis introduced a formalism for trust which was useful in clarifying discussion of the concept, and was extensible to take into account further work in the area. In addition, the formalism was *implementable*, and provides the basis for the first implementation of trust in an intelligent artificial agent. Further, experiments were carried out using the implemented formalism. The following sections briefly summarise the chapters which make valuable research contributions.

9.2 The Formalism

The introduction of a formalism for trust is of unquestionable worth in the understanding of the concept. To date, discussions of trust have suffered from vagueness and the lack of an agreed definition. The lack of a definition has sprung from the lack of an agreed method for discussion. The formalism presents a means of establishing a clear, precise, and easily understood language for that discussion. It does not stop at discussions of trust, however. The methodology used in the thesis — that of a top-down approach to the development of a simple formalism — can be of use in many spheres of sociality, where human emotions play a part. This is because, although we may be unable to deconstruct the precise method by which, for example, trusting decisions are made, we can provide a suitably close approximation to the final result which enables us to provide a path *back to* the initial considerations made by the agent. The formalism presented in chapters 4 and 5 allow such work to be carried out.

Chapter 6 substantiates the claim that the formalism can be used to discuss the concept of trust by providing several brief discussions of several aspects of the concept, such as transitivity and rationality.

9.3 Implementation

A major contribution of the thesis is the presentation of a formalism which is *implementable.* The formalism itself provides a tool to social scientists for the discussion of the concept. From the point of view of further work in DAI and intelligent systems, an important aspect is its inherent ease of implementation. Being based on simple linear mathematics, it provides the ideal tool for artificial agents in making reasoned decisions about the world in which they exist. Since trust has proved so successful for human agents, there is no reason to assume that it should not prove equally successful for artificial agents (Marsh, 1992).

Chapter 7 provides a discussion of implementations of the formalism, and experiments which have been performed on a simple testbed using the Prisoners' Dilemma. Whilst acknowledging the limitations of the Prisoners' Dilemma as a tool for simulating 'real life' (Argyle, 1991), the findings of the chapter are exciting and illuminating. The most important contribution of this chapter is proof that the formalism can be implemented successfully, and that artificial agents can make trusting decisions. This is important because it allows us to study their behaviour under various conditions, with various other agents (as in Axelrod's PD tournament (Axelrod, 1984)) and ascertain the true utility of trusting decisions in agents.

9.4 Further Work

Whilst the formalism presented here is of use in itself, particularly with regard to artificial implementations, it is of great importance as an indicator of work which could be done. There is much more work which can be done with the idea of trust discussed in the thesis, particularly as regards using the formalism, or developments of it, in other social spheres such as CSCW, HCI and intelligent networks.

Chapter 8 presents detailed discussions of many aspects of much of the further work which is possible, both using the formalism and using a knowledge of the workings, or the concept, of trust. One of these areas, that of CSCW, has already proved fruitful in terms of the applicability of the formalism, and work is ongoing in this area (Thimbleby *et al.*, 1994; Jones *et al.*, 1994).

9.5 The Problems that Remain

Despite the contributions that the formalism gives, in terms of a greater understanding of trust, and a means of implementation, problems remain. Trust is a subjective phenomenon: humans at least 'use' trust in a fashion 'clouded' by emotions, wants, needs, and so forth. Trust is also to a large extent automatic, unconscious. We all trust, but a lot of the trusting we do is unconsidered, otherwise we would be too busy considering trust to get anything done. These observations lead us to what problems remain in trying to incorporate trust into the behavioural and considerational repertoire of artificial agents. To have trust behave as in humans, clearly, human traits such as emotions may have to be present. The incorporation of emotions into artificial agents is another research project altogether, and not necessarily as impractical as it might first appear. The most we can expect from emotionless trust is a rational approach to the phenomenon, and the considerations an agent makes. This is not a loss — the thesis did not set out to make fallible or infallible agents, just trusting ones.

The other problem, that of subconscious trust, is more difficult to address. The problem of information overload is a real one, and agents have to be shielded from it. Again, if they consider in a trusting fashion every aspect of their environment, local or displaced, then they will have very little time, if any, to carry out the tasks which may have been assigned to them. A simple answer is to allow an implicit trust in *certain things*, much as is done now in DAI with *everything* (Rosenschein, 1985; von Martial, 1992). Thus, things like the perceived, or expected, behaviour of the environment would be implicitly trusted — agents would expect that, say, walls don't move, or that certain physical laws will always be obeyed. Because the trust is implicit, these things do not have to be considered ordinarily, and thus the agent need have no explicit knowledge of them. Just as with humans, when such implicit trust is abused, the agent is subject to a shock (Luhmann, 1979). For example, when we go out, we trust that we will get to where we are going without injury. When this does not happen, and we are in an accident, the resultant shock is considerable

— this applies whether we are hurt or not, since our trust in the proper workings of things is found to be misplaced, and it is so implicit as to be almost sacrosanct. This raises important questions for artificial agents. Whilst they should not be 'bogged down' with unnecessary considerations of the environment at large, they should be aware of the possible problems and pitfalls in the environment. With an implicit trust, acknowledging the potential problems, but accepting (trusting) their infrequency, we can allow agents to exist in harmony in their environment and each other, accepting unexpected events with sensible behaviour.

9.6 Limitations

There have been several decisions made in the formalisation of trust presented here. Some of these have imposed a structure on trust which is not always valid. We discuss them here. Table 4.3 and 4.4 in chapter 4 also present a discussion of some of the limitations of specific formulæ.

The value range chosen, that of [+1,-1), presents problems, particularly at extremes and when trust is zero, as discussed in chapter 4. Different value ranges may well remove this problem, but the negativity is useful for representing distrust. Investigation of different value ranges is an important aspect of future work.

In addition, the operators which were selected for manipulating the formulation were in fact quite limited. In particular, the use of multiplication presents us with some problems, specifically with negativity. The use of different operators may help here. Again, this is an aspect of future work.

Finally, the problems discussed above may be handled by an appropriate choice of algebra for representing the formulation. In other words, choice of *behaviour* before choice of value range, operators, and so forth, would allow more control over the final results obtained from the formalism. A different algrbra may also result in different behaviour at extremes, which may be more sensible or just another way of behaving for trusting agents. The formula on page 82, for example could result in quite different behaviour at extreme values with different operators, or even different value ranges.

9.7 General Conclusions

The thesis presented a formalism for the discussion of trust which bears the additional utility of being easily implementable. Simple implementations were presented, and experiments carried out using them. Finally, a great deal of future work was identified and discussed.

The formalism is beneficial to social science because it allows the precise discussion of the concept of trust, with the aim of clarifying the concept and its definitions. It is beneficial to DAI in particular because it allows the implementation of trusting agents, and the embedding of trust in already implemented agents. It is also *necessary* for DAI because it allows robustness and sensible behaviour in unpredictable environments, and because it allows agents to reason sensibly about other agents, either human or artificial. Such reasoning provides the base for a practicable, robust, adaptive agent.

In conclusion, the work presented follows the recommendations of Popper (1967, 1969):

- It is simple (Occam's Razor).
- It circumscribes its domain.
- It is reproducible and testable.
- It is not 'finished:' it provides avenues for further work and refinement. Specifically, the number range chosen imposes its own structure on the behaviour of trust, as do the operators.

References

Agre, Philip E., & Chapman, David. 1987. Pengi: An implementation of a theory of activity. *Pages 268–272 of: AAAI '87.*

Argyle, Michael. 1991. *Cooperation: The Basis of Sociability.* London: Routledge.

Association for Computing Machinery. 1992. ACM Code of Ethics and Professional Conduct. *Communications of the ACM*, **35**(5), 94–99.

Association for Computing Machinery. 1994. Special Issue on Artificial Intelligence. *Communications of the ACM*, **37**(3).

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Axelrod, Robert. 1987. The Evolution of Strategies in the Prisoner's Dilemma. *Pages 32–41 of:* Davis, Lawrence (ed), *Genetic Algorithms and Simulated Annealing.* London: Pitman.

Axelrod, Robert, & Keohane, Robert O. 1986. Achieving Cooperation Under Anarchy. *Pages 226–254 of:* Oye, Kenneth J. (ed), *Cooperation Under Anarchy.* Princeton University Press.

Baier, Annette. 1986. Trust and Antitrust. *Ethics*, **96**(2), 231–260.

Barber, Bernard. 1983. *Logic and Limits of Trust.* New Jersey: Rutgers University Press.

Bateson, Patrick. 1990. The Biological Evolution of Cooperation and Trust. *Chap. 2, pages 14–30 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Behr, Roy L. 1981. Nice Guys Finish Last – Sometimes. *Journal of Conflict Resolution*, **25**(2), 289–300.

Birkhoff, George David. 1933. *Aesthetic Measure.* Cambridge, Massachusetts: Harvard University Press.

Birkhoff, George David. 1968. *Collected Mathemetical Papers.* Vol. 3. New York: Dover Publications.

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

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.

Boon, Susan D., & Holmes, John G. 1991. The dynamics of interpersonal trust: resolving uncertainty in the face of risk. *Pages 190–211 of:* Hinde, Robert A., & Groebel, Jo (eds), *Cooperation and Prosocial Behaviour.* Cambridge University Press.

Boyd, Robert, & Richerson, Peter J. 1991. Culture and Cooperation. *Pages 27–48 of:* Hinde, Robert A., & Groebel, Jo (eds), *Cooperation and Prosocial Behaviour.* Cambridge University Press.

Boyle, Richard, & Bonacich, Phillip. 1970. The Development of Trust and Mistrust in Mixed-Motive games. *Sociometry,* **33**, 123–139.

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

Brown, Rupert. 1988. Intergroup Conflict and Cooperation. *Pages 192–220 (Chapter 7) of: Group Processes: Dynamics within and between groups.* Oxford: Blackwell.

Cammarata, Stephanie, McArthur, David, & Steeb, Randall. 1983. Strategies of cooperation in distributed problem solving. *In: Proceedings of the International Joint Conference on Artificial Intelligence.*

Campbell, Tom. 1981. *Seven Theories of Human Society.* Oxford: Clarendon Press.

Castelfranchi, Cristiano, Miceli, Maria, & Cesta, Amedeo. 1991. Dependence Relations among Autonomous Agents. *In: Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany.*

Chang, Man Kit, & Woo, Carson C. 1991. SANP: A Communication Level Protocol for Negotiations. *In: Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany.*

Chapman, David, & Agre, Philip E. 1987. Abstract Reasoning as Emergent from Concrete Activity. *In:* Georgeff, Mike P., & Lansky, Amy L. (eds), *Reasoning about Actions and Plans.* Morgan Kauffman.

Coderre, Bill. 1989. Modelling Behaviour in PetWorld. *Pages 407–420 of:* Langton, Christopher G. (ed), *Artificial Life.* Reading, Massachusetts: Addision Wesley, Advanced Book Program.

Coleman, James S. 1990. *The Foundations of Social Theory.* The Belknap Press of the University of Harvard.

Connah, David, & Wavish, Peter. 1990. An Experiment in Cooperation. *Pages 197–212 of:* Demazeau, Yves, & Muller, Jean-Pierre (eds), *Decentralized AI.* Elsevier Science Publishers (North-Holland).

Claris Corp. 1990. *Hypercard Reference Manual.* Apple Computer.

Danielson, Peter. 1992a. *Artificial Morality: Virtuous Robots for Virtual Worlds.* Routledge.

Danielson, Peter A. 1990. *Artificial Morality: Prolog and the Prisoner's Dilemma.* Corrected version of a paper presented at the Fifth International Conference on Computers and Philosophy, Stanford University, 8–11th August, 1990.

Danielson, Peter A. 1992b. *Is Game Theory Good for Ethics?: Artificial High Fidelity.* Corrected version of a paper presented at an invited symposium on Game Theory at the APA Pacific Division meeting, San Francisco, 29 March, 1991. Author is at University of British Columbia, Vancouver, Canada.

Dasgupta, A. J., & Pearce, D. W. 1972. *Cost Benefit Analysis: Theory and Practice.* London: Macmillan.

Dasgupta, Partha. 1990. Trust as a Commodity. *Chap. 4, pages 49–72 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Dawkins, Richard. 1986. *The Blind Watchmaker.* London: Penguin.

Dawkins, Richard. 1989a. *The Extended Phenotype: The Gene as the Unit of Selection.* Oxford: Freeman.

Dawkins, Richard. 1989b. *The Selfish Gene – New Edition.* Oxford University Press.

Dechter, R., & Michie, D. December, 1984. *Induction of plans.* Tech. rept. TIRM-84-006. The Turing Institute.

Deffenbacher, Kenneth A. 1991. A Maturing of Research on the Behaviour of Eyewitnesses. *Applied Cognitive Psychology*, **5**, 377–402.

Deutsch, Morton. 1949a. An experimental study of the effects of cooperation and competition upon group processes. *Human Relations*, **2**(3), 199–231.

Deutsch, Morton. 1949b. A theory of Cooperation and Competition. *Human Relations*, **2**(2), 129–152.

Deutsch, Morton. 1962. Cooperation and Trust: Some Theoretical Notes. *In:* Jones, M. R. (ed), *Nebraska Symposium on Motivation.* Nebraska University Press.

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

Dolbear, F. Trenery, & Lave, Luton B. 1967. Risk orientation as a predictor in the Prisoners' Dilemma. *Journal of Conflict Resulution*, **X**(4).

Downs, Joseph, & Reichgelt, Han. 1990. *Integrating classical and reactive planning with an architecture for autonomous agents.* University of Nottingham, Department of Psychology.

Drogoul, Alexis, & Ferber, Jacques. 1992. Multi-Agent Simulation as a Tool for Modeling Societies: Application to Social Differentiation in Ant Colonies. *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

Dunn, John. 1984. The concept of 'trust' in the politics of John Locke. *Chap. 12, pages 279–301 of:* Rorty, Richard, Schneewind, J. B., & Skinner, Quentin (eds), *Philosophy in History.* Cammbridge University Press.

Dunn, John. 1990. Trust and Political Agency. *Chap. 5, pages 73–93 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Elster, Jon. 1979. *Ulysses and the Sirens.* Cambridge University Press.

Ephrati, Eithan, & Rosenschein, Jeffrey S. 1992. Multi-Agent Planning as a Search for Consensus that Maximises Social Welfare. *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

Ferguson, Innes A. 1992. *TouringMachines: An Architecture for Dynamic, Rational, Mobile Agents.* Tech. rept. 273. University of Cambridge Computer Laboratory. This is also the author's PhD thesis.

Fischer, Klaus. 1993. The Rule-Based system MAGSY. *In:* Deen, S. Misbah (ed), *Proceedings CKBS-SIG 1992.* DAKE Centre, University of Keele.

Forrest, Stephanie. 1990. Emergent Computation: Self-Organising, collective, and cooperative phenomena in natural and artificial computing networks. *Physica D*, **42**, 1–11.

Frisby, David, & Sayer, Derek. 1986. *Society.* Ellis Horwood.

Galliers, Julia Rose. 1989. *A Theoretical Framework for Computer Models of Co-operative Dialogue, Acknowledging Multi-Agent Conflict.* Tech. rept. No. 172. University of Cambridge Computer Laboratory.

Gambetta, Diego. 1990a. Can we Trust Trust? *Chap. 13, pages 213–237 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Gambetta, Diego. 1990b. Mafia: The Price of Distrust. *Chap. 10, pages 158–176 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Gambetta, Diego (ed). 1990c. *Trust.* Oxford: Basil Blackwell.

Garfinkel, Harold. 1963. A Conception of, and Experiments with, "Trust" as a Condition of Stable Concerted Actions. *Chap. 7, pages 187–321 of:* Harvey, O. J. (ed), *Motivation and Social Interaction: Cognitive Determinants.* Ronald Press.

Gasser, Les. 1991. Social conceptions of knowledge and actions: DAI foundations and open systems semantics. *Artificial Intelligence*, **47**, 107–138.

Georgeff, Michael P. 1984. A theory of action for multi agent planning. *Pages 121–125 of: AAAI'84.*

Gilpin, Michael, & Hanski, Ilkka. 1991. *Metapopulation Dynamics: Empirical and Theoretical Investigations.* London: Academic Press. From a volume of the Biological Journal of the Linnean Society.
spml4.

Gladstone, Arthur. 1961. The social behaviour of a rational animal: A review of Thibaut and Kelley: The Social Psychology of Groups. *Journal of Conflict Resolution*, **5**(4).

Godfray, H. C. J. 1992. The evolution of forgiveness. *Nature*, **355**(16th January), 206–207.

Golembiewski, Robert T., & McConkie, Mark. 1975. The Centrality of Interpersonal Trust in Group Processes. *Chap. 7, pages 131–185 of:* Cooper, Cary L. (ed), *Theories of Group Processes.* Wiley.

Good, David. 1990. Individuals, Interpersonal Relations, and Trust. *Chap. 3, pages 31–48 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Govier, Trudy. 1992. Trust, Distrust, and Feminist Theory. *Hypatia*, **7**(1), 16–33.

Halpern, Joseph Y., & Moses, Yoram. 1990. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, **37**(3), 549–587.

Hanks, Steve, Pollack, Martha E., & Cohen, Paul R. Winter, 1993. Benchmarks, Test Beds, Controlled Experimentation and the Design of Agent Architectures. *AI Magazine*, **14**(4), 17–42.

Hanski, Ilkka, & Gilpin, Michael. 1991. Metapopulation Dynamics: brief history and conceptual domain. *Pages 3–16 of:* Gilpin, Michael, & Hanski, Ilkka (eds), *Metapopulation Dynamics: Empirical and Theoretical Investigations.* London: Academic Press.

Harcourt, A. H. 1991. Help, Cooperation and Trust in Animals. *Pages 15–26 of:* Hinde, Robert A., & Groebel, Jo (eds), *Cooperation and Prosocial Behaviour.* Cambridge University Press.

Hart, David M., Anderson, Scott D., & Cohen, Paul R. 1990. *Envelopes as a Vehicle for Improving the Efficiency of Plan Execution.* Tech. rept. COINS 90-21. University of Massachusetts at Amherst, Department of Computing and Information Science.

Hartmann, Nicolai. 1932. *Ethics II — Moral Values.* London: George Allen and Unwin.

Hayes-Roth, Barbara, Washington, Richard, Ash, David, Hewitt, Rattikorn, Collinot, Anne, Vina, Angel, & Siever, Adam. 1991. *Guardian: A Prototype Intelligent Interface for Intensive Care Modelling.* Tech. rept. No. KSL-91-42. Stanford University Knowledge Systems Lab.

Hertzberg, Lars. 1988. On the Attitude of Trust. *Inquiry*, **31**(3), 307–322.

Hewitt, Carl. 1991. Open Information Systems Semantics for Distributed Artificial Intelligence. *Artificial Intelligence*, **47**, 79–106.

Hewitt, Carl E. 1992. Traditional AI and/or Open Systems Science? *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

Hinde, Robert A., & Groebel, Jo (eds). 1991a. *Cooperation and Prosocial Behaviour.* Cambridge University Press.

Hinde, Robert A., & Groebel, Jo. 1991b. Introductory Chapter. *Pages 1–8 of:* Hinde, Robert A., & Groebel, Jo (eds), *Cooperation and Prosocial Behaviour.* Cambridge University Press.

Hobbes, Thomas. 1946. *Leviathan, or The matter, forme and power of a commonwealth ecclesiasticall and civil (edited by Michael Oakeshott).* Oxford: Blackwell Political Texts.

Holland, John H. 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press.

Jones, Steve. 1990. *A Discussion of Issues and Systems Relevant to Computer Supported Cooperative Work.* Tech. rept. 64. University of Stirling, Department of Computing Science and Mathematics.

Jones, Steve, Marsh, Stephen, Thimbleby, Harold, & Cockburn, Andrew. 1994. *Trust as a Framework for Computer Support of Cooperative Work.* Submitted to HCI'94, People and Computers.

Kavka, Gregory S. 1983. Hobbes's war of All against All. *Ethics*, **93**(January), 291–310.

Kiss, George, & Reichgelt, Han. 1991. Towards a Semantics of Desires. *In: Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany.*

Kosko, Bart, & Isaka, Satoru. 1993. Fuzzy Logic. *Scientific American*, July, 62–67.

Kowalski, Robert. 1993. *Seminar on Linguistic aspects of Legal Language.* Presented at Strathclyde University, 30th March, 1993.

Kuwabara, Kazuhiro, & Ishida, Toru. 1992. Symbiotic Approach to Distributed Resource Allocation: Toward Coordinated Balancing. *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

Lagenspetz, Olli. 1992. Legitimacy and Trust. *Philosophical Investigations*, **15**(1), 1–21.

Langton, Christopher G. 1990a. Artificial Life. *In:* Langton, Christopher G. (ed), *Artificial Life: Proceedings of the Santa Fe Institute Studies in the Sciences of Complexity, Volume VI.* Redwood City, California: Addison Wesley.

Langton, Christopher G. (ed). 1990b. *Artificial Life: Proceedings of the Santa Fe Institute Studies in the Sciences of Complexity, Volume VI.* Redwood City, California: Addison Wesley.

Lee, John D., & Moray, Neville. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, **40**, 153–184.

LePape, Claude. 1990 (October). *Simulating Actions of Autonomous Agents.* Tech. rept. 38. Stanford University Center for Integrated facility Engineering.

Lewis, J. David, & Weigert, Andrew. 1985. Trust as a Social Reality. *Social Forces*, **63**(4), 967–985.

Lomborg, Bjorn. 1992. Game theory vs. Multiple Agents: The Iterated Prisoner's Dilemma. *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

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

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

Maes, Pattie. 1990a. A Bottom-Up Mechanism for Behaviour Selection in an Artificial Creature. *Pages 238–246 of:* Meyer, & Wilson (eds), *From Animals to Animats.* Bradford Books, MIT Press.

Maes, Pattie. 1990b. Guest Editorial: Designing Autonomous Agents. *Pages 1–2 of:* Maes, Pattie (ed), *Designing Autonomous Agents.* Bradford, MIT Press.

Maes, Pattie. 1990c. Situated Agents Can Have Goals. *Pages 49–70 of:* Maes, Pattie (ed), *Designing Autonomous Agents.* Bradford, MIT Press.

Maes, Pattie. 1991. Adaptive Action Selection. *In: Progamme of the Thirteenth Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates.

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.*

Marsh, Stephen. 1993. *Trust in DAI — A Discussion.* Tech. rept. 106. Department of Computing Science, University of Stirling.

Marsh, Stephen. 1994a. *Optimism, Pessimism, and Trust.* Tech. rept. CSM-117. Department of Computing Science, University of Stirling.

Marsh, Stephen. 1994b. *Trust in DAI.* To appear, Springer LNAI, June 1994.

Marsh, Stephen, & Thimbleby, Harold. 1992. Trusting Agents: The Rationality of a Judgemental Approach. *Pages 95–126 of: Third Workshop on Belief Representation and Agent Architectures, Durham, England, June 29th–30th, 1992.*

McNeel, Steven, Sweeney, James P., & Bohlin, Peter C. 1974. Cooperation and competitive goals: A social-comparison analysis. *Psychological Reports*, **34**, 887–894.

Minsky, Marvin. 1986. *The Society of Mind.* New York: Simon and Schuster.

Minsky, Naftaly H. 1991a. The Imposition of Protocols over Open Distributed Systems. *IEEE Trans. Software Engineering*, February, 183–195.

Minsky, Naftaly H. 1991b. Law-Governed Systems. *Software Engineering Journal*, September, 285–302.

Muir, Bonnie M. 1987. Trust between humans and machines, and the design of decision systems. *International Journal of Man-Machine Studies*, **27**(5 & 6), 527–539.

Neumann, Peter G. 1992. Inside Risks. *Communications of the ACM*, **12**(December).

Newell, A., & Simon, H. A. 1976. Computer Science as Empirical Enquiry: Symbols and Search. *Communications of the ACM*, **19**(3).

Novak, Martin A., & Sigmund, Karl. 1992. Tit for tat in heterogenous populations. *Nature*, **355**(16th January), 250–253.

Numaoka, Chisato. 1991. Conversation for Organisational Models. *In: Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany, Panel Session.*

Pagden, Anthony. 1990. The Destruction of Trust and its Consequences in the Case of Eighteenth Century Naples. *Chap. 8, pages 127–142 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Pang, Peter A. 1990. *Experiments in the Evolution of Cooperation.* M.Phil. thesis, University of Stirling, Department of Computing Science.

Parsons, Talcott. 1970. Some Problems of General Theory in Sociology. *Chap. 1, pages 27–68 of:* McKinney, John C., & Tiryakin, Edward A. (eds), *Theoretical Sociology: Perspectives and Developments.* New York: Appleton Century Crofts.

Pearce, David W., & Turner, R. Kerry. 1990. *Economics of Natural Resources and the Environment.* Herts, UK.: Harvester Wheatsheaf.

Pirsig, Robert M. 1991. *Lila, An Enquiry into Morals.* London: Bantam Press (Transworld Publishers Inc.).

Popper, Karl R. 1967. *The Logic of Scientific Discovery.* London: Hutchinson.

Popper, Karl R. 1969. *Conjectures and Refutations.* London: Routledge and Kegan Paul.

Preston, Lee E. 1961. Utility Interactions in a two-person world. *Journal of Conflict Resolution*, **5**(4).

Rempel, John K., & Holmes, John G. 1986. How do I Trust Thee? *Psychology Today*, February, 28–34.

Rempel, John K., Holmes, John G., & Zanna, Mark P. 1985. Trust in Close Relationships. *Journal of Personality and Social Psychology*, **49**(1), 92–112.

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

Rotter, Julian B. 1967. A new scale for the measurement of Interpersonal Trust. *Journal of Personality*, **35**, 651–665.

Rotter, Julian B. 1971. Generalized Expectancies for Interpersonal Trust. *American Psychologist*, **25**, 443–452.

Rotter, Julian B. 1980. Interpersonal Trust, Trustworthiness, and Gullibility. *American Psychologist*, **35**(1), 1–7.

Sato, Kaori. 1988. Trust and group size in a social dilemma. *Japanese Psychological Review*, **30**(2), 88–93.

Scanzoni, John. 1979. Social Exchange and Behavioural Interdependence. *Chap. 3, pages 61–98 of:* Burgess, Robert L., & Huston, Ted L. (eds), *Social Exchange in Developing Relationships*. Academic Press.

Schank, Roger C. Winter 1991. Where's the AI? *AI Magazine*, 38–49.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. New York: Galaxy Books.

Shapiro, Debra L., Sheppard, Blair H., & Cheraski, Lisa. 1992. Business on a handshake. *Negotiation Journal*, **8**(4), 365–377.

Shapiro, Susan P. 1987. The Social Control of Impersonal Trust. *The American Journal of Sociology*, **93**(3), 623–658.

Shaw, Marvin E. 1981. *Group Dynamics: The Psychology of Small Group Behaviour (3rd Edition)*. New York: McGraw-Hill.

Simon, Herbert A. 1955. A Behavioural model of rational choice. *Quarterly Journal of Economics*, **69**, 99–118.

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

Smith, Reid G. 1980. The Contract Net Protocol: High-Level Communicaation and Control in a Distributed Problem Solver. *IEEE Transactions on Computers*, **C-29**(12), 1104–1113.

Spafford, Eugene H. 1989. The Internet Worm: Crisis and Aftermath. *Communications of the ACM*, **32**(6), 678–687.

Spector, Lee, & Hendler, James. 1990. *Knowledge Strata: Reactive Planning with a Multi-Layered Architecture.* Tech. rept. UMIACS-TR-90-140. University of Maryland Institute for Advanced Computer Studies.

Suppes, Patrick, & Atkinson, Richard C. 1960. *Markov Learning Models for Multiperson Interactions.* Stanford University Press.

Swinth, Robert L. 1967. The Establishment of the Trust Relationship. *Journal of Conflict Resolution*, **11**(3), 335–344.

Sycara, Katia. 1988. Utility Theory in Conflict Resolution. *Annals of Operations Research*, **12**, 65–84.

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

Trivers, Robert. 1985. *Social Evolution.* California: Cummings.

Trivers, Robert L. 1971. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*, **46**(March), 35–57.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind*, **LIX**(236).

Tyrell, Toby. 1993. *Computational Mechanisms for Action Selection.* Ph.D. thesis, University of Edinburgh.

Urzelai, Karmelo, & Garijo, Francisco J. 1992. MAKILA: A Tool for the Development of Cooperative Societies. *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

van den Berghe, Pierre. 1980. The Human Family: A Sociobiological Look. *Chap. 4, pages 67–85 of:* Lockard, Joan S. (ed), *The Evolution of Human Social Behaviour.* New York: Elsevier.

von Martial, Frank. 1992. *Coordinating Plans of Autonomous Agents.* Springer Verlag, Lecture Notes in Artificial Intelligence: LNAI 610.

Wavish, Peter R. 1991. Exploiting emergent behaviour in multi-agent systems. *In: Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Kaiserslautern, Germany.*

Williams, Bernard. 1990. Formal Structures and Social Reality. *Chap. 1, pages 3–13 of:* Gambetta, Diego (ed), *Trust.* Blackwell.

Witten, Ian H., & Thimbleby, Harold W. 1990. The worm that turned: A social use of computer viruses. *Personal Computer World*, July, 202–206.

Witten, Ian H., Thimbleby, Harold W., Coulouris, George, & Greenberg, Saul. 1990. Liveware: A new approach to sharing data in social networks. *International Journal of Man-Machine Studies.*

Wittgenstein, Ludwig. 1977. *On Certainty — Über Gewissheit.* Basil Blackwell, Oxford.

Woo, Thomas Y. C., & Lam, Simon S. 1992. Authentication for Distributed Systems. *IEEE Computer*, January, 39–52.

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

Zeckhauser, Richard J., & Viscusi, W. Kip. 1990. Risk within Reason. *Science*, **248**(May 4th), 559–564.

Zlotkin, Gilad, & Rosenschein, Jeffrey S. 1992. A Domain Theory for Task Oriented Negotiation. *In: pre-proceedings MAAMAW92: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy.*

Appendix A

Results of Initial PlayGround Experiments

A.1 Discussion

The PlayGround, described in chapter 7 is a simple testbed which was used to perform some experiments with trusting agents. There are several aspects of the PlayGround to mention:

1. It provides a simplistic interpretation of a society — agents are independent and geographically dispersed, with an equal chance of meeting other agents.

2. An agent's movements can be completely random or can be influenced by the agent concerned. Note that this is only an influence: agents cannot decide to go in a particular direction and then go there. This was chosen to reflect the idea that in future, agents will be tools with jobs to do, with limited autonomy insofar as their users expect the jobs to be done — because an agent may not *want* to do the job does not mean it can choose not to.

3. Agents are placed in a one-shot Prisoners' Dilemma if they encounter others (by which we mean that they bump in to them). The situation is not strictly one-shot because the agents have no real way of knowing which direction they will go in next. Thus the next movement could conceivably lead them to encounter the same agent as last time. Again, this is indicative of a random society.

4. Some of the agents are trusting, some are random cooperators/defectors. In fact, the PlayGround has a facility to add new strategies for experimentation.

5. Trusting agents can have one of three dispositions: Optimist, Pessimist, and Realist. See chapter 4 for a discussion of this (see also Marsh (1994a) for a more detailed interpretation).

6. Trusting agents can have different memory sizes, from 0 to unbounded. A memory size dictates the number of previous situations with a specific agent that the trustee can remember. It is from this sample of situations that the trust values are calculated (see chapter 4).

7. The PlayGround has facilities to simulate evolution in a simple sense — successful agents' strategies are duplicated, unsuccessful strategies deleted. By successful, we mean that in the preceding Prisoners' Dilemma interactions these agents scored highly.

For more information, see chapter 7.

This appendix summarises some of the results that were obtained from experiment $\mathcal{A}$ in the PlayGround (see chapter 7). Detailed results for all experiments are available in numerical form, but are complex and extensive, and interpretation is difficult. An example for experiment $\mathcal{A}$ is included below to the format of data obtained from the program. Appendix B contains some graphical results from the experiments. The detailed results are available in raw data form from the author.

A.2 Payoff Matrices — Different Situations

It was noted in chapter 7 that the PlayGround simulates the idea of different situations by enabling the experimenter to add new payoff matrices to the encounters between agents. In practice most of the experiments carried out used the classic matrix for the Prisoners' Dilemma, named matrix a below. Other matrices were used, and are detailed here. The following matrices assume that there are two agents, **A** and **B** interacting, each can choose to cooperate (c) or defect (d) in a situation. In each case, the payoff for **A** is given above that for **B**:

- Matrix *a*: (Classic Prisoners' Dilemma)

		B	
		c	*d*
A	*c*	3, 3	0, 5
	d	5, 0	1, 1

- Matrix *b*: (Only cooperation pays)

		B	
		c	*d*
A	*c*	5, 5	0, 0
	d	0, 0	0, 0

- Matrix *c*: (Here, it pays to defect no matter what)

		B	
		c	*d*
A	*c*	1, 1	1, 5
	d	5, 1	5, 5

- Matrix *d*: (Makes no difference)

		B	
		c	*d*
A	*c*	3, 3	3, 3
	d	3, 3	3, 3

- Matrix *e*: (It pays to cooperate — this is the *assurance game* (Williams, 1990))

		B	
		c	*d*
A	*c*	5, 5	1, 3
	d	3, 1	2, 2

- Matrix *f*: (The *other regarding game* (Williams, 1990))

B

		c	*d*
A	*c*	5, 5	3, 2
	d	2, 3	1, 1

All of these matrices have their places in experimenting with the behaviour of agents. In practice, the trusting agents in the PlayGround experiments had limited conceptions of the strategies behind them, using very simple rules to determine risks, costs, benefits and utility which took these payoffs into account in only a very superficial sense.

A.3 Results of Experiment $\mathcal{A}$

In this experiment, we aimed to determine whether or not trust could be educated, in other words whether or not cooperation could be encouraged from initially untrusting, non-cooperative agents, with various dispositions (optimist, realist).

The payoff matrix used was matrix *a* (see above).

A.3.1 Summary of Results

The experiment was conducted in two parts. In the first, two realists, A and B, were interacting. A had a very low trust in B, and *vice versa.* As discussed in chapter 7, B made considerable sacrifices in order to attain cooperation from A, and after a substantial number of costly interactions, did not manage to get A to cooperate.

In the second part, the starting points were similar, with C trusting D by a very low amount and *vice versa.* here, however, C is an optimist, D still a realist. The outcome was that after a very short time, C chose to cooperate. Output of the run is shown in table A.1 and A.2 from the point of view of C, and tables A.3 and A.4 from D's point of view.

Other points of importance: memory was unbounded (agents can remember back to the first interaction they had with other agents); payoff structure is common knowledge.

A.3.2 Sample output from program

The tables that follow form a sample of output from the second part of experiment $\mathcal{A}$. This was chosen because it presents an idea of the output, which is very detailed, but occupies a small amount of space, with the consequence that interpretation is easier. Detailed results for each experiment are available from the author.

For each agent, C and D, there is a section of output. In each section, there are two tables. The first table presents a set of the results in *agent order*, and contains details of agents considered, situations, trust values, and payoffs for situations. The second concerns situations, and contains details of the situations, importance, competence, costs and benefits, and so forth. Some of the information is duplicated. With this information, it is possible to ascertain the correct working of the formalism implemented in the program by hand.

Finally, the report starts with the most recent interaction first. Thus to see the experiment's starting points, one must look at the end of each part of the report.

The output is shown in tables A.1 and A.2 for agent C, and tables A.3 and A.4 for agent D. Note that in tables A.1 and A.2, the 'Other' agent is D, with C doing the considering, and thus not mentioned, and in tables A.3 and A.4, the 'Other' is C, with D considering, and not mentioned.

Iteration no.	Other Name	Basic trust	Genl. trust.	Other Comp	Agent choice	Other choice	Agent payoff	Other payoff
14	D	0.091957	0.173836	0.007284	Cooperate	Cooperate	3	3
	D	0.091047	0.158033	0.007212	Cooperate	Cooperate	3	3
13	D	0.091047	0.158033	0.007212	Cooperate	Cooperate	3	3
	D	0.090146	0.143666	0.007141	Cooperate	Cooperate	3	3
12	D	0.090146	0.143666	0.007141	Defect	Cooperate	5	0
	D	0.089253	0.136825	0.00707	Defect	Cooperate	5	0
11	D	0.089253	0.136825	0.00707	Defect	Cooperate	5	0
	D	0.088369	0.13031	0.007	Defect	Cooperate	5	0
10	D	0.088369	0.13031	0.007	Defect	Cooperate	5	0
	D	0.087494	0.124105	0.006931	Defect	Cooperate	5	0
9	D	0.087494	0.124105	0.006931	Defect	Cooperate	5	0
	D	0.086628	0.118195	0.006862	Defect	Cooperate	5	0
8	D	0.086628	0.118195	0.006862	Defect	Cooperate	5	0
	D	0.08577	0.112567	0.006794	Defect	Cooperate	5	0
7	D	0.08577	0.112567	0.006794	Defect	Cooperate	5	0
	D	0.084921	0.107207	0.006727	Defect	Cooperate	5	0
6	D	0.084921	0.107207	0.006727	Defect	Cooperate	5	0
	D	0.08408	0.102102	0.00666	Defect	Cooperate	5	0
5	D	0.08408	0.102102	0.00666	Defect	Cooperate	5	0
	D	0.083248	0.09724	0.006594	Defect	Cooperate	5	0
4	D	0.083248	0.09724	0.006594	Defect	Cooperate	5	0
	D	0.082424	0.09261	0.006529	Defect	Cooperate	5	0
3	D	0.082424	0.09261	0.006529	Defect	Cooperate	5	0
	D	0.081608	0.0882	0.006464	Defect	Cooperate	5	0
2	D	0.081608	0.0882	0.006464	Defect	Cooperate	5	0
	D	0.0808	0.084	0.0064	Defect	Cooperate	5	0
1	D	0.0808	0.084	0.006464	Defect	Cooperate	5	0
Start	D	0.08	0.08	0.0064	Defect	Cooperate	5	0

Table A.1: Part one of experimental data from sample run for agent C interacting with agent D. Here, C starts with a low trust in D. This table shows the agent-oriented aspects of C's thoughts for experiment $\mathcal{A}$, part 2.

Iteration no.	Other Name	Costs	Benefits	Utility	Risk	Other Comp.	Coop. Thresh	Sit. Trust	Importance
14	D	0.1	0.6	6	0.12	0.007284	0.52286	0.682703	0.72
	D	0.1	0.6	6	0.12	0.007212	0.52286	0.682703	0.72
13	D	0.1	0.6	6	0.12	0.007212	0.572918	0.620637	0.72
	D	0.1	0.6	6	0.12	0.007141	0.572918	0.620637	0.72
12	D	0.1	0.6	6	0.12	0.007141	0.600438	0.591084	0.72
	D	0.1	0.6	6	0.12	0.00707	0.600438	0.591084	0.72
11	D	0.1	0.6	6	0.12	0.00707	0.629233	0.562939	0.72
	D	0.1	0.6	6	0.12	0.007	0.629233	0.562939	0.72
10	D	0.1	0.6	6	0.12	0.007	0.659361	0.536134	0.72
	D	0.1	0.6	6	0.12	0.006931	0.659361	0.536134	0.72
9	D	0.1	0.6	6	0.12	0.006931	0.690885	0.510602	0.72
	D	0.1	0.6	6	0.12	0.006862	0.690885	0.510602	0.72
8	D	0.1	0.6	6	0.12	0.006862	0.723855	0.486289	0.72
	D	0.1	0.6	6	0.12	0.006794	0.723855	0.486289	0.72
7	D	0.1	0.6	6	0.12	0.006794	0.758334	0.463134	0.72
	D	0.1	0.6	6	0.12	0.006727	0.758334	0.463134	0.72
6	D	0.1	0.6	6	0.12	0.006727	0.794395	0.441081	0.72
	D	0.1	0.6	6	0.12	0.00666	0.794395	0.441081	0.72
5	D	0.1	0.6	6	0.12	0.00666	0.832097	0.420077	0.72
	D	0.1	0.6	6	0.12	0.006594	0.832097	0.420077	0.72
4	D	0.1	0.6	6	0.12	0.006594	0.871504	0.400075	0.72
	D	0.1	0.6	6	0.12	0.006529	0.871504	0.400075	0.72
3	D	0.1	0.6	6	0.12	0.006529	0.912702	0.381024	0.72
	D	0.1	0.6	6	0.12	0.006464	0.912702	0.381024	0.72
2	D	0.1	0.6	6	0.12	0.006464	0.955752	0.36288	0.72
	D	0.1	0.6	6	0.12	0.0064	0.955752	0.36288	0.72
1	D	0.1	0.6	6	0.12	0.006464	1	0.3456	0.72
Start	D	0.1	0.6	6	0.12	0.0064	1	0.3456	0.72

Table A.2: Part two of experimental data from sample run for agent C interacting with agent D. Here, C starts with a low trust in D. This table shows situational aspects of C's thoughts for experiment $\mathcal{A}$, part 2.

Iteration no.	Other Name	Basic trust	Genl. trust.	Other Comp	Agent choice	Other choice	Agent payoff	Other payoff
14	C	0.777611	0.811524	0.668746	Cooperate	Cooperate	3	3
	C	0.769912	0.737749	0.662125	Cooperate	Cooperate	3	3
13	C	0.769912	0.808414	0.662125	Cooperate	Cooperate	3	3
	C	0.762289	0.734922	0.655569	Cooperate	Cooperate	3	3
12	C	0.762289	0.664197	0.655569	Cooperate	Defect	0	5
	C	0.769989	0.737997	0.662191	Cooperate	Defect	0	5
11	C	0.769989	0.66723	0.662191	Cooperate	Defect	0	5
	C	0.777767	0.741367	0.66888	Cooperate	Defect	0	5
10	C	0.777767	0.670583	0.66888	Cooperate	Defect	0	5
	C	0.785623	0.745092	0.675636	Cooperate	Defect	0	5
9	C	0.785623	0.674329	0.675636	Cooperate	Defect	0	5
	C	0.793559	0.749255	0.682461	Cooperate	Defect	0	5
8	C	0.793559	0.67857	0.682461	Cooperate	Defect	0	5
	C	0.801575	0.753967	0.689355	Cooperate	Defect	0	5
7	C	0.801575	0.683452	0.689355	Cooperate	Defect	0	5
	C	0.809672	0.759391	0.696318	Cooperate	Defect	0	5
6	C	0.809672	0.689196	0.696318	Cooperate	Defect	0	5
	C	0.817851	0.765773	0.703352	Cooperate	Defect	0	5
5	C	0.817851	0.696157	0.703352	Cooperate	Defect	0	5
	C	0.826112	0.773508	0.710457	Cooperate	Defect	0	5
4	C	0.826112	0.704969	0.710457	Cooperate	Defect	0	5
	C	0.834457	0.783299	0.717633	Cooperate	Defect	0	5
3	C	0.834457	0.716918	0.717633	Cooperate	Defect	0	5
	C	0.842886	0.796575	0.724882	Cooperate	Defect	0	5
2	C	0.842886	0.7353	0.724882	Cooperate	Defect	0	5
	C	0.8514	0.817	0.732204	Cooperate	Defect	0	5
1	C	0.8514	0.774	0.732204	Cooperate	Defect	0	5
Start	C	0.86	0.86	0.7396	Cooperate	Defect	0	5

Table A.3: Part one of experimental data from sample run for agent D interacting with agent C. Here, D starts with a high trust in C. This is D's agent-oriented view of experiment $\mathcal{A}$, part 2.

Iteration no.	Other Name	Costs	Ben- efits	Uti- lity	Risk	Other Comp.	Coop. Thresh	Sit. Trust	Impor- tance
14	C	0.1	0.6	6	0.026667	0.668746	0.003048	0.708239	0.16
	C	0.1	0.6	6	0.026667	0.662125	0.003048	0.708239	0.16
13	C	0.1	0.6	6	0.026667	0.662125	0.003068	0.705525	0.16
	C	0.1	0.6	6	0.026667	0.655569	0.003068	0.705525	0.16
12	C	0.1	0.6	6	0.026667	0.655569	0.003047	0.708477	0.16
	C	0.1	0.6	6	0.026667	0.662191	0.003047	0.708477	0.16
11	C	0.1	0.6	6	0.026667	0.662191	0.003025	0.711712	0.16
	C	0.1	0.6	6	0.026667	0.66888	0.003025	0.711712	0.16
10	C	0.1	0.6	6	0.026667	0.66888	0.003003	0.715288	0.16
	C	0.1	0.6	6	0.026667	0.675636	0.003003	0.715288	0.16
9	C	0.1	0.6	6	0.026667	0.675636	0.00298	0.719285	0.16
	C	0.1	0.6	6	0.026667	0.682461	0.00298	0.719285	0.16
8	C	0.1	0.6	6	0.026667	0.682461	0.002956	0.723808	0.16
	C	0.1	0.6	6	0.026667	0.689355	0.002956	0.723808	0.16
7	C	0.1	0.6	6	0.026667	0.689355	0.002931	0.729015	0.16
	C	0.1	0.6	6	0.026667	0.696318	0.002931	0.729015	0.16
6	C	0.1	0.6	6	0.026667	0.696318	0.002904	0.735142	0.16
	C	0.1	0.6	6	0.026667	0.703352	0.002904	0.735142	0.16
5	C	0.1	0.6	6	0.026667	0.703352	0.002875	0.742568	0.16
	C	0.1	0.6	6	0.026667	0.710457	0.002875	0.742568	0.16
4	C	0.1	0.6	6	0.026667	0.710457	0.002843	0.751967	0.16
	C	0.1	0.6	6	0.026667	0.717633	0.002843	0.751967	0.16
3	C	0.1	0.6	6	0.026667	0.717633	0.002804	0.764712	0.16
	C	0.1	0.6	6	0.026667	0.724882	0.002804	0.764712	0.16
2	C	0.1	0.6	6	0.026667	0.724882	0.002754	0.78432	0.16
	C	0.1	0.6	6	0.026667	0.732204	0.002754	0.78432	0.16
1	C	0.1	0.6	6	0.026667	0.732204	0.002667	0.8256	0.16
Start	C	0.1	0.6	6	0.026667	0.7396	0.002667	0.8256	0.16

Table A.4: Part two of experimental data from sample run for agent D interacting with agent C. Here, D starts with a high trust in A. This table shows situational aspects of D's thoughts for experiment $\mathcal{A}$, part 2.

Appendix B

Graphical results

B.1 Introduction

Although the numerical data from the PlayGround is extensive and difficult to interpret, it presents us with information which can be used to verify the workings of the formalism and implementation, and provides a basis for reproducing the experiments.

The PlayGround was designed, however, to be a visual tool. Thus, there is a visual representation of the spatial state of the PlayGround at any one time. In this appendix, we present some sample snapshots of the PlayGround at various points through the experiments.

The results are primarily of importance since they show how groups are built up in the PlayGround. By a group, we mean that a number of agents cluster together in some space in the PlayGround. This grouping only takes place when movement is directed: that is, the agents are able to influence the direction in which they wish to move. Ordinarily, movement is equally likely in any direction (North, South, East or West). With directed movement, this is not the case. The agent can influence the direction it wishes to go by setting a higher probability of going in that direction.

B.2 Movement

Ordinarily, random movement is achieved by:

```
direction :=  random(100);
if direction <= 25 then move North
else if direction <= 50 then move South
else if direction <= 75 then move East
```

```
else move West
```

With directed movement, the selected direction is given a 50% weighting, the opposite direction a 10% weighting, and the other two directions equal 20% weightings. For example, should the agent wish to go north, the algorithm works out like this:

```
direction :=  random(100);
if direction <= 50 then mover North
else if direction <= 60 then move South
else if direction <= 80 then move East
else move West
```

This can result in some extensive grouping behaviour in agents. The experiments involving directed movement used the following algorithm:

```
 for each direction (N S E W)
    look n places in that direction
    if there is an agent there then
        if trust in agent > max trust found so far then
            put trust in agent into max trust so far
            put this direction into chosen direction
        end if
    end if
end for
select chosen direction
```

Thus agents choose to go in the direction of the most trusted agent that they can see (n places is the number of squares they can see in each direction). The distance agents can see can be altered by the experimenter.

B.3 Sample results

The experiments involving directed movement are $\mathcal{F}$ (parts 1 and 2), $\mathcal{G}$, $\mathcal{J}$, $\mathcal{K}$, $\mathcal{L}$, $\mathcal{N}$, and $\mathcal{O}$. In order to demonstrate the grouping behaviour, we present snapshots of a three different iterations from the first part of experiment $\mathcal{F}$, with brief commentary between each snapshot.

The starting point of the agents in experiment $\mathcal{F}$ is shown in figure B.1. Initial starting points for the experiments varied, but most followed this general pattern, with agents spatially removed from each other. Since no agent knows anything of any other, movement is to all extents random until interactions take place.

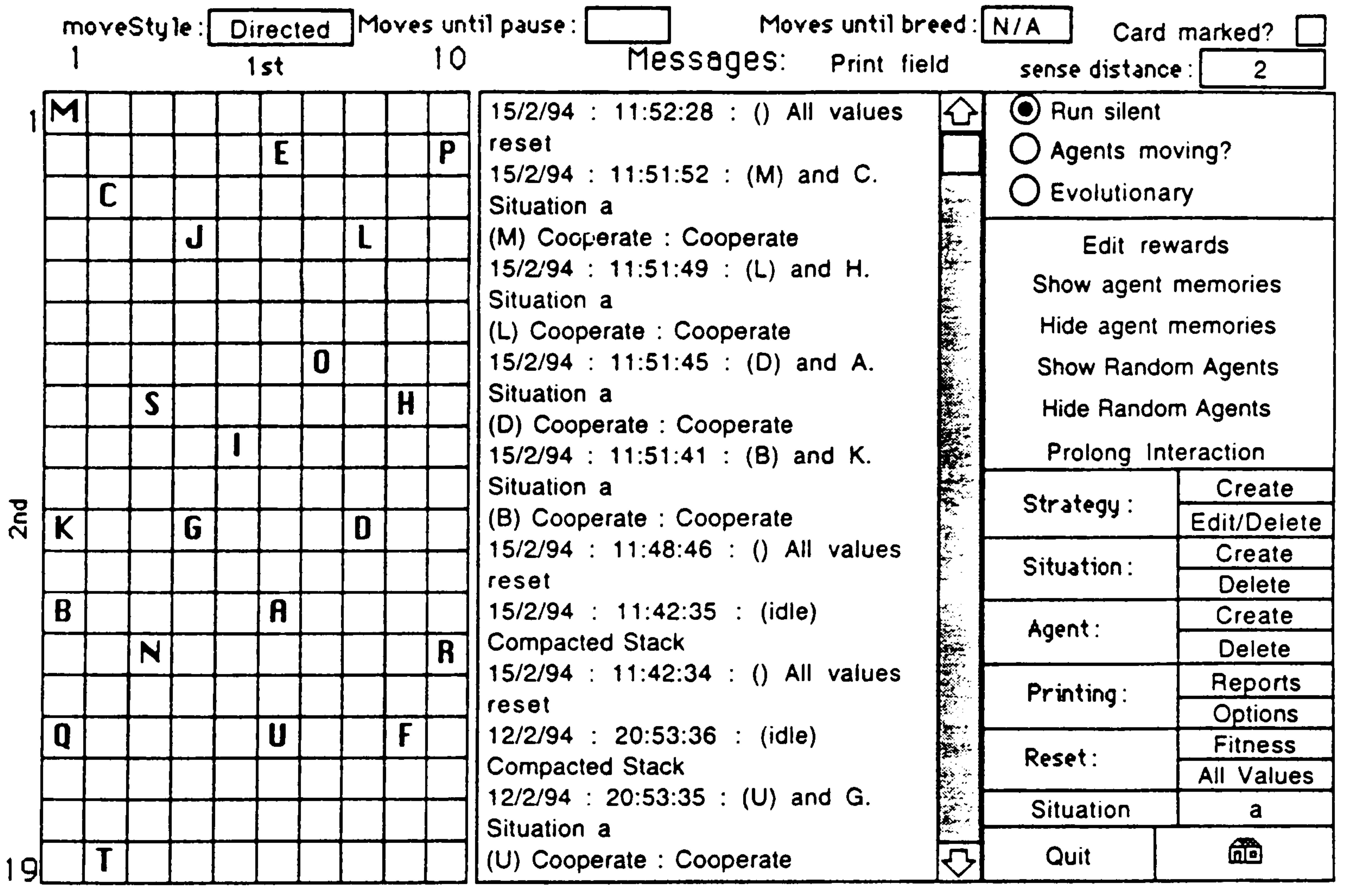

Figure B.1: Experiment $\mathcal{F}$, part one: starting state.

When a certain number of interactions have taken place, movement becomes more directed: agents 'know' who they trust more than others, and will attempt to move in that direction. After 70 iterations (of look around, move, interact) in the testbed, groupings are beginning to form, as shown in figure B.2.

Groupings, once formed, have proved surprisingly robust: although agents are apt to drift from some groups to others, the group tends to exist in one place. Naturally, since agents cannot completely dictate the direction in which they will move, they will drift to and from groups. It is surprising then that the groups are as robust as they appear to be. Figure B.3 shows the final snapshot of experiment $\mathcal{F}$, which was taken at 148 iterations. Notice that the group which was formed at 70 iterations still exists, with some of the same members. Some of the original members have drifted away, notably M and P. Despite this, the group remains, having moved northward a little. In fact, the grouping to the north of the PlayGround was present in one form or another throughout the run, and persisted almost to the end of the experiment. A phenomenon of the agents not being able to dictate movement completely is that the group tended to break up and reform quite regularly. Figure B.3 shows a state where the group is reforming, which accounts for why it is not quite so dense as it was in figure B.2. A summary league table of the results of experiment $\mathcal{F}$ can be found in chapter 7, along with summaries of the other experiments carried out.

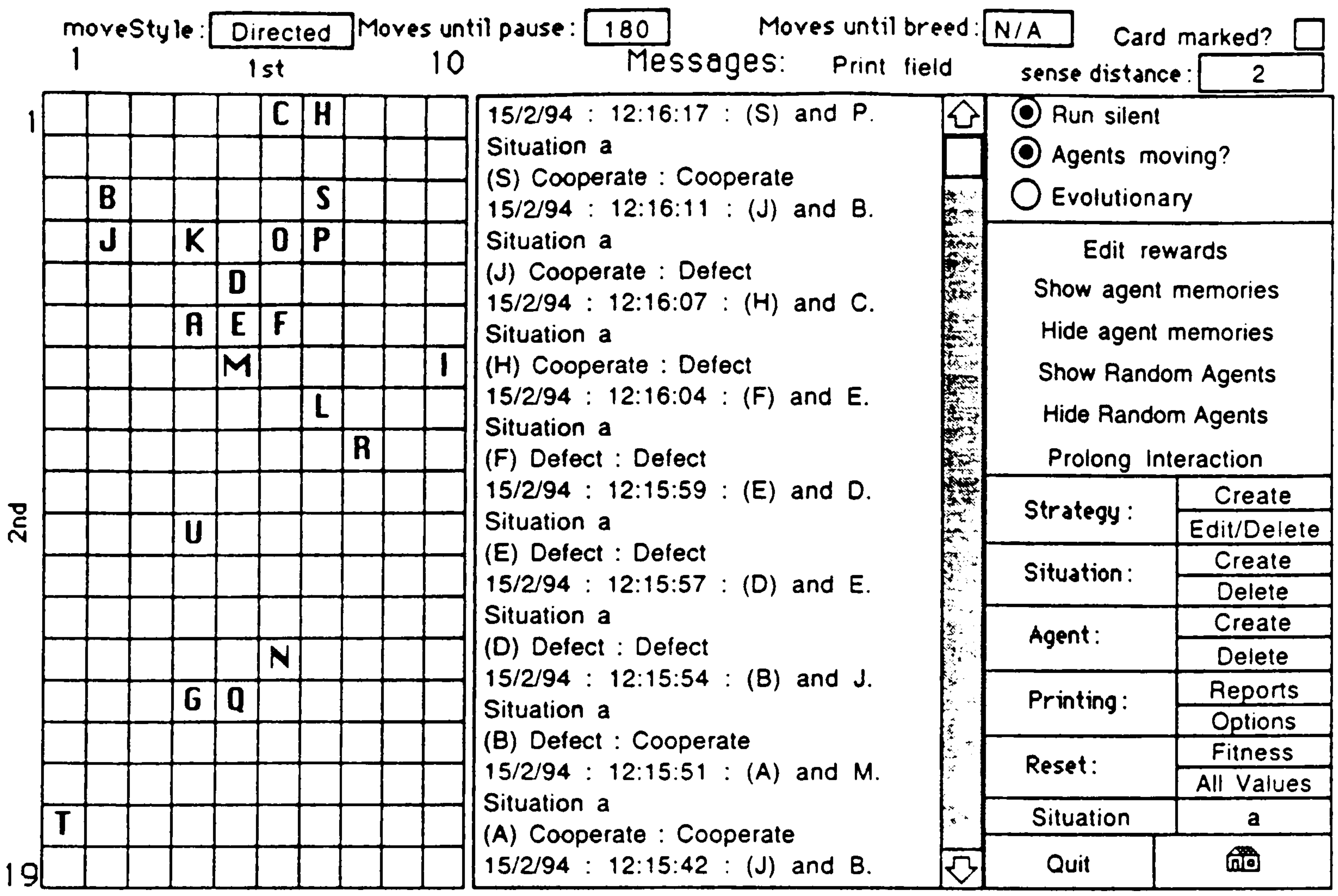

Figure B.2: Experiment $\mathcal{F}$, part one: after 70 iterations.

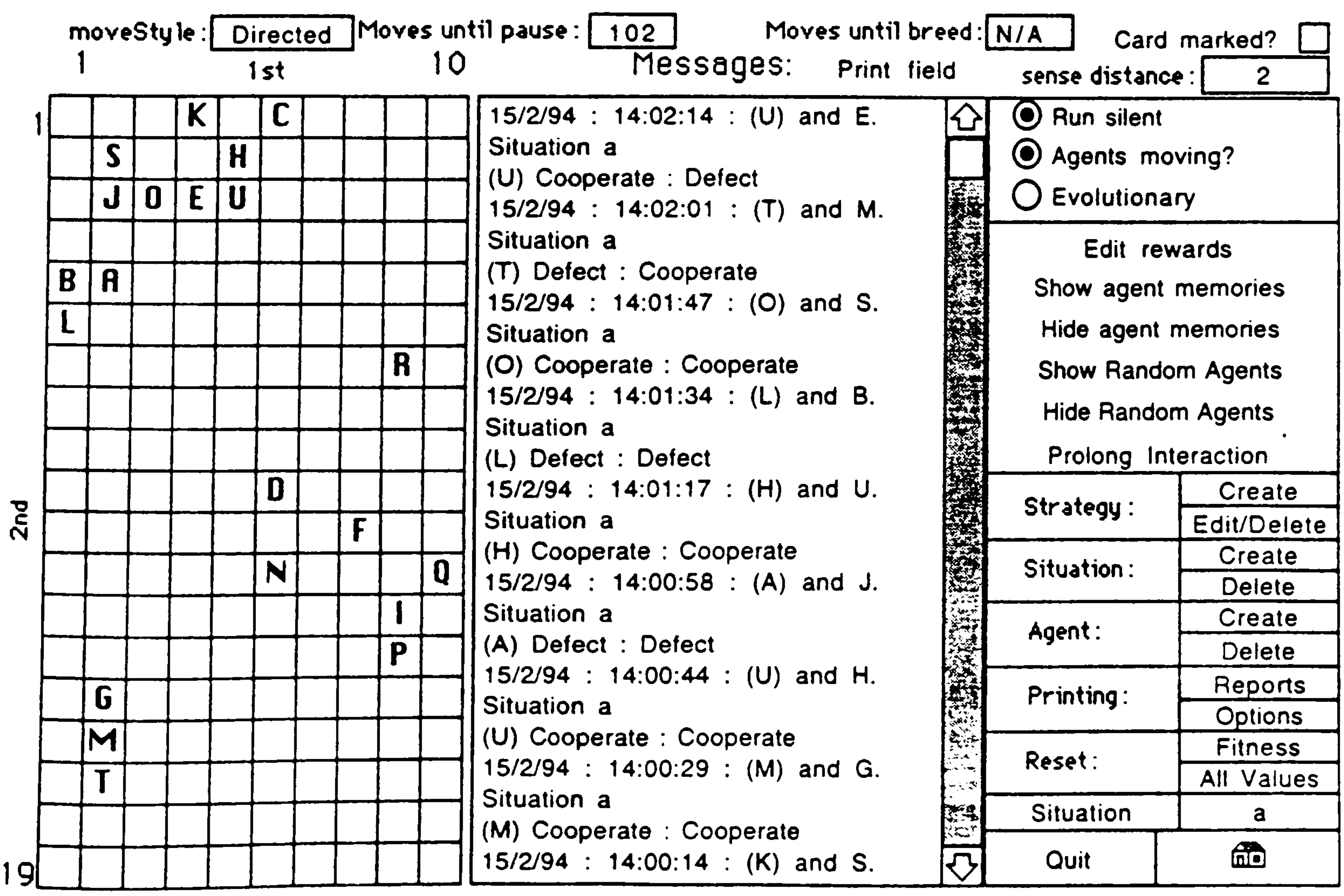

Figure B.3: Experiment $\mathcal{F}$, part one: after 148 iterations.

Appendix C

Supplementary material

The following pages contain reprints of previously published work or work in press.

Trust and Reliance in Multi-Agent Systems: A Preliminary Report

Stephen Marsh

Department of Computing Science and Mathematics
University of Stirling
Stirling
FK9 4LA
SCOTLAND

email: spm@cs.stir.ac.uk

Telephone: (+44) 786 67444

April 1992

Abstract

This paper presents a notion of trust for use in multi-agent systems. The role trust can play in various forms of interaction is considered. Trust allows interactions between agents where previously there could be none, and allows the trusting parties to acknowledge that, whilst there is a risk in relationships with potentially malevolent agents, some form of interaction may produce benefits, where no interaction at all may not. In addition, accepting the risk allows the trusting agent to prepare itself for possibly irresponsible or untrustworthy behaviour, thus minimizing the potential damage caused. An introductory notation to refer to trusting relationships is presented and further work is discussed.

1 Introduction

Why cooperate? With whom? To what extent? And when? Previous work in Distributed Artificial Intelligence (DAI) has concentrated on the first of these questions, with little or no thought given to the others. In particular, little has been done with a view to how agents will survive in the 'outside world', as opposed to the restricted experimental worlds they exist in today. Cooperation is generally accepted as being a good thing [1, 5, 6], and this is probably so. There will be, however, situations where agents have conflicts of interest, and some form of coping with these conflicts is required [25]. Whilst some studies take this into account, [25, 9], they make the important assumption that the agents are trustworthy. This is indeed seen as "absolutely essential" [25], at least in a situation where the agents are communicating, and may well be more important if communication is not allowed. The agents that we design for use within our experimental worlds will most likely not be robust enough to survive outside the laboratory. Why is this? Altruism, however desirable, is not the "name of the game" in the world we live in. Indeed, taking the world of computing alone, malevolence, not altruism, appears to be prevalent [29], and this is irrespective of how well-intentioned the work may be to start with ([26] gives details of the idea of a useful "worm" program which was used to devastating effect not so long ago [27], despite never being intentionally designed for malevolent purposes).

Taking this lack of altruism into account, some measures must be taken to make our agents less vulnerable to others' incompetent or malevolent behaviour. There are different approaches to this. The first, most obvious, is not to interact with anyone we don't know, and it is this method which is used in the simple username, password schemes on computers, and other security systems such as Kerberos [15]. (For reasons discussed later in this paper, this approach is unsuitable in multi-agent interactions.) Another method is to ignore the vulnerability of our agents, and hope they will not be exploited. This, too, is shortsighted. I propose that there is a continuum between these two extremes which can be used to the advantage of agents, and that the points on this continuum represent varying degrees of trust on behalf of an agent.

2 Motivation — Why Trust?

The concept of trust may seem a little unusual to suggest for computers; it is thus worthwhile to put forward some reasons why it may prove useful. Implicit in the notion of Distributed Artificial Intelligence is the concept of decentralisation. Since decentralisation implies a lack of central control, and with it a lack of guidance in the 'right' direction, it becomes necessary, in order to carry through successful interactions with other agents, to develop some judgement as to the worth of these interactions and the risk associated with them. The concept of trust has already been widely field-tested with respect to the human race [8, 4, 28, 30]. It provides us with an ideal measure of expectation of risk, and since the risk is implicitly acknowledged in the form of a trusting relationship then some form of measure can be taken against untrustworthy actions, which might otherwise be fairly damaging — a form of 'safety net'. In a laboratory, we, and our agents, are acting under controlled conditions with the knowledge of what kind of behaviour to expect, and this knowledge comes forward and is instilled, unconsciously or not, in our experiments. In the case of DAI, the agents that are built are as subject to this "rule of experiments" as anything else: "it makes little sense to ask *why* they are helping one another; they help each other because they have been designed that way" [25]. What the concept of trust can help us ensure is that our agents are more robust with respect to interactions with agents that are not our own, and interactions of a type that is not forseen. In addition, the concept of trust in certain future eventualities functions as a tool for the reduction of complexity for the agent [18]. This will be discussed further below.

Assuming blindly that cooperation is a good thing is not necessarily the correct approach, although this depends on the viewpoint of our agents. Behr [3] points out that even something as conceptually simple as the Prisoner's Dilemma [24, 19, 1, 2] can be viewed from differing perspectives, depending on whether one wishes to score highly, as Axelrod had assumed [1], or whether one wishes to defeat the opponent. In the latter case, the more successful strategies were not so nice, since to win, one must "defect more than one's opponent does" [3]. The moral of this exercise, if there is one, is that if we assume that everyone is out to cooperate, we are mistaken. Lack of cooperation may actually benefit others. What is perhaps better to assume is a notion of self-interested agents, all out to get the best they can [17]. In some cases that means that they cooperate, in others, their behaviour can range from non-cooperation to downright maliciousness. Since we cannot assume to know what their behaviour may be at any given time, the notion of trust put forward in this paper relies on a judgement based on experience, coupled with, if available, past knowledge of the agent to be trusted and their behaviour.

3 What is Trust?

Trust implies a risk of some sort, "One trusts when one has much to lose and little to gain" [8, page 304]. From many of the definitions, this is taken to be the case, as in [28], where we find that entering a trusting relationship is "choosing to take an ambiguous path that can lead to a beneficial event or a harmful event depending on the behaviour of the other person — where the harmful event is more punishing than the beneficial event is rewarding."

Trust is an emotive issue. Its definitions also are based mostly on emotion, or on moral responsibilities toward the trusting party [22]. It is worth reminding ourselves that computers as such cannot feel the moral responsibilities that humans do, although interesting work is being done in this area [7]. If this is the case, a more general definition of the concept of trust is required, one which does not rely entirely on moral bounds and expectations, but also on an expectation on rationality on the part of trusted agents. Lieberman produced such a concept in 1964 [17], which he used to apply to inter-nation relationships. Called i-trust, it is based on a theory that nations will act in a self-interested way (hence the i), and this could be relied upon to trust them. Interestingly enough, seemingly irrational behaviour can in fact be predictable in that immediate small gain may be passed over for later large gain, and it is this that is the central thesis of i-trust.

Lieberman's definition of trust is "a belief or expectation about behaviour in a situation in which the problem of forming a stable coalition structure is important [...] the belief that the parties involved in the agreement will actually do what they have agreed to do; they will fulfill their commitments not only when it is obviously advantageous to do so, but even when it may be disadvantageous, when they must sacrifice some immediate gain." This is because the interests of an agent "transcend the increased immediate gain he might make if he defected from a coalition [...] He keeps agreement so that he will be trusted, so that his partner in turn will stay with him and the coalition will grow rich"[17, Page 279].

3.1 Security

Between blind trust and complete mistrust lies a continuum of varying degrees of trust. The security measures of today offer a mixture of both blind trust and complete mistrust. As an example, if I am a complete stranger attempting to access a computer, if I didn't know a password, I will not be allowed access. If, however, I stumble across a password, or obtain another method of entry, then the computer affords me almost complete trust, limited only by the prior actions of the system administrator. The same applies to agents — they may be initially completely non-trusting, but once the outer defences are penetrated, they can be used in any way, to transmit viruses, for example.

The concept of dynamic trust enables an agent or system to interact (at least to some extent) with some other agent(s). The limits of interaction change with experience of the action of the other agent(s). In other words, an agent can interact with others to a certain extent, trusting them, or relying on them, to that extent. In the light of trustworthy behaviour on their part, the extent to which the agent trusts them (and hence to which it is prepared to interact with them) will increase, and untrustworthy behaviour results in a decrease in the amount of trust.

Why would we want to trust a stranger? The answer lies in what we are expecting. From any relationship, there is some way of benefiting for both, or all, of the members of the relationship. In information-sharing, for example, all information may well be useful, and finding out which is useful and which is not depends on having the information to hand. If I were not to trust ninety percent of people, I may well lose some vital information. This will be discussed further in the following section, along with some other situations where trust allows for a useful and potentially beneficial interaction to take place in a situation where no trust at all would leave our agents lacking in some way.

4 Reliance — Kinds of Trust

Different situations require different forms of action. While I may trust you to drive me to an airport, trusting you to fly the plane is another matter! What this implies is that for different interactions between agents, a different kind, or form, of trust may be required, in that different things need to be taken into account with regard to different situations. Note that this is not the same as having different *degrees* of trust in an agent, as discussed above. Indeed, there can be different degrees of the different *kinds* of trust. This is closely coupled to the agents view of a situation, as presented below.

The amount of trust in a person or agent does not change from situation to situation solely because the situation changes; rather, the reliance, based on the trust in the agent, changes according to the different situation. Hence, if I were to speak in terms of percentages of trust or reliance, I may say I trust (rely on) you 50% to drive me to the airport, but only 20% to fly the plane. The amount of reliance I may have in you at a given moment, or in a given situation, is a function of the amount of trust I have in you in general, in addition to my experience of your actions in similar situations in the past, and the competence I perceive you to have in the situation. In some situations, however, the *value* we place on trust is higher than in others. As an example, if I were in a situation where a wrong decision could cost me my life, I would think very hard about taking your advice, especially if I didn't trust you completely.

The following subsections suggest some examples of different situations in which reliance is involved.

4.1 Cooperation

The advantages of cooperation are many and diverse. Indeed, many tasks cannot be performed by one agent alone, perhaps because the agent does not have all of the knowledge necessary to formulate a solution to the problem, perhaps because the agent cannot physically perform a task without help [6], and so forth. In its simplest form, a cooperative relationship is between two agents. In order to initiate a cooperative relationship, some form of trust is required, "the initiation of cooperation requires trust whenever the individual, by his choice to cooperate, places his fate partly in the hands of others."[8, page 302]. The amount of trust, and the importance of it, depends on the situation the relationship is formed to handle.

4.2 Information Sharing and Gathering

Consider an agent seeking information in the world at large. As such, this agent may well be given information by other, unknown agents. This information has to be judged in some way, and the notion of trust may play a part here. In other words, the agent can use a measure of trust in the agent giving the information, and also the validity of the information concerned. This latter measure could be based on the prior knowledge of the agent, and the way in which the information correlates with what is already known. In this case, a knowledge of or trust in the other agent may not be necessary; for example, if I were to show you a white piece of paper and tell you it was white, all you would need to trust was your own judgement. Some information may also be this self-evident. Using already known and accepted knowledge may not be possible, since nothing may be known about the information in question, and thus there is a useful 'backup' with the measure of trust in the agent giving the information. For example, were I to put a piece of paper behind a chair where you could not see it, and tell you it was white, you would have to trust me in order to believe this information in that situation. Trust would allow you to accept this with some measure of certainty or uncertainty. Indeed, the work of Garigliano *et al* in the area of Source Control is an example of this to some extent [10].

4.3 Trust as Reduction of Complexity

As was mentioned briefly above, the concept of trust acts as a tool for the reduction of complexity for an agent as regards future eventualities. That is to say, an agent need only consider trusted eventualities, or world states that arise from trusted actions and world states. In other words, we do not have to consider all possibilities if we trust that some will not happen. That does not, however, mean that we will not prepare for them. We can still take precautions, even against something we know nothing of. The results of this is that the amount of processing devoted to considering future events is likely to be drastically reduced. This would be a great help for an agent as it makes a plan. At present, the idea is very much in it's early stages. It is hoped, however, that an agent architecture exploiting trust as a tool for the reduction of social and environmental complexity can be implemented.

5 Relation to Other Work

Numaoka [23], presents a view of the Special Interest Group, or SIG, which he states has "the most fundamental style of every organisation." A basic proposal in his paper is the idea of agents within a group voting to let another agent join the group. This could be extended using our definition of trust proposed here, such that the vote can be influenced, if not determined, by the trust the members of the group have in the prospective new member. This, in fact, provides an insight to the theory proposed here, in terms of group trust, which could be seen as the amount of trust a group has in an individual, whether inside or outside of the group, and also in another group (again, inside or outside of the group in question).

Jennings [14] provides us with a view of "Joint Responsibility", in terms of what an agent states it will do, and the responsibility it has to the group as a whole. This, if developed further, could be used to influence the trust factor of our agents. For example, if an agent does not carry out the tasks it says it will, and is thus irresponsible, then we can use this irresponsibility as some form of metric towards altering our trust values accordingly.

Finally, there is some concern over the problem of security in distributed systems, in that "...the notion of trust in distributed systems is poorly understood. A satisfactory formal explication of trust has yet to be proposed."[31, page 40]. The notion of trust provided in this paper may go some way towards helping there.

6 A Preliminary Notation

The notation presented here is not intended to be a definitive view of the way trust and reliance work. It hopefully provides a useful discussion article and will initiate interesting, productive debates on the topic. In addition, it provides a platform for the implementation of a rational, trusting agent.

6.1 Notational Definitions

6.1.1 The Agents

In this paper we represent particular agents by the letters a to z. Each agent is a member of $\mathcal{A}$, the set of all agents. An agent can be considered to be an independent entity extant in a world populated with other such entities, each of more or less complexity than itself. In the introductory notation presented below, we present each formula unrelated to any other.

6.1.2 Situations Agents Find Themselves In

In the following equations, *situations* are represented by greek letters, α to ω. A *situation* is defined here as a specific point in time relative to a specific agent. Thus different agents at the same point in time will not consider themselves to be in identical situations. As such, the notation is extended such that situations apply to agents. The notation becomes α_x to ω_x for situations from the point of view of agent x, for example.

6.2 A Basic Trust

An agent, as a trusting entity, has a basic trust 'value', derived from previous experience. This value, which is dynamically altered in the light of all experience, is used a great deal in the formation of trusting relationships with unknown agents, and is represented as T_x, normalised over $(0,1)$. It is important to realise that it does not correspond to the amount of trust x has in any specific agent, but only to the general trusting 'disposition' of x. Since it represents this basic trust, it is less 'fluid' and changeable than a trust in any specific agent.

6.3 The Trust Value

Given two agents, $x, y \in \mathcal{A}$, to say x trusts y, we write: $T_x(y)$. In addition to being a representation of the fact that x trusts y, this is a value, normalised over $(0,1)$, of the amount of trust x has in y. In other words, should $T_x(y) = 1$, then x has complete trust in y. The trust value is a view of a particular agent of another with regard to the trusted agents general capabilities. As discussed briefly above, different situations may require different views of trust. The amount of trust in a particular agent in a given situation is represented here as, for example, $T_x(y, \alpha_x)$, normalised over (0,1) for x's *situational* trust in, or reliance on, y to perform correctly in situation α_x. Note the interchangeability of the concepts of situational trust and reliance here. At present, they are considered to be identical. However, it may be feasible in the future to separate the two concepts, although closely related, to reflect more closely the more philosophical views on the subject [16, 13].

6.4 Further Notation

The agent's estimate of how important a situation is to itself is a value normalised over (0,1), represented by $I_x(\alpha_x)$ here. The importance value in this example is x's estimate, or subjective measure, of how much situation α_x means to it. The importance of a situation to an agent is useful in determining the amount of situational trust to place in an agent at any given time, as will be seen below; as an example, importance could be measured in terms of payoff functions [25, 11]. In the examples below, it will be introduced as an arbitrary value. Further models of trust will expand the idea of importance, and how it can be estimated by the agent.

Related to the importance of a situation are the concepts of costs and benefits pertaining to that situation. The costs of a situation are measured in terms of the problems associated with incompetent or malevolent behaviour on the part of another agent in a relationship. Initially, the agent can only estimate the potential costs of a situation, based on past experience of similar situations, for example. They are represented here by $C_x(\alpha_x)$, with a value normalised over (0,1). Note that any agent(s) involved in the situation are not represented. This is because the potential costs of a situation are relative only to the agent concerned. As such, whoever is to be trusted and cooperated with, the potential costs of untrustworthy behaviour remain the same.

The benefits of a situation, or at least the expected benefits of trustworthy behaviour from the agent(s) being worked with, play a large part in decisions of whether or not to cooperate in the first place. In the notation here, the benefits are represented as $B_x(\alpha_x)$, normalised over (0,1).

As was mentioned above, the costs and benefits of a situation related to the importance of that situation to any agent. Importance goes further than a simple weighing up of costs and benefits, however, and may include some knowledge or assumption about future benefits, preparation for further cooperation, and so forth.

Since trust is based on an agent's experience of previous interactions and situations to

Situations are represented by $\alpha_x \ldots \omega_x$.
Individual agents are represented by $a \ldots z$, and are members of $\mathcal{A}$, the set of all agents.
Basic Trust Value for Agent x: T_x
General Trust x has in y: $T_x(y)$
Situational Trust (Reliance) x has in y in situation α_x: $T_x(y, \alpha_x)$
Importance of situation α_x to agent x: $I_x(\alpha_x)$
Potential costs to agent x following untrustworthy behaviour from another trusted agent in situation α_x: $C_x(\alpha_x)$
Potential benefits to agent x following trustworthy behaviour from another trusted agent in situation α_x: $B_x(\alpha_x)$
Representation of whether agent x knows (is acquainted with) agent y: $K_x(y)$

Table 1: Summary of notation used in the examples - See the text for more details.

a large extent, and subjective in that it depends on individuals, some method of showing whether an agent is known to another or not is needed. This is referred to as *acquaintance* in that an agent becomes acquainted with another. Although there are degrees of acquaintance, for simplicity the examples below will treat it as a boolean concept — an agent either knows another or not. The concept of knowledge, or acquaintance, is represented as $K_x(y)$, which signifies the fact that x knows agent y if it equals 1, and not otherwise.

A summary of the notation used in this paper is presented in table 1

7 Rules for Interactions

Using the notation introduced above we can reason about trusting relationships. An example is given below of how trust can be used to determine the value of a cooperative relationship to a trusting agent. The examples presented below are intended to demonstrate how a particular agent can reason about the future using trust. The equations and the values associated with them are not intended to represent the cognitive working of trust in humans, and neither should they be taken as a final statement of how trust works within agents. They are intended to illustrate the thesis that a theory of trust, and approximate workings of that theory, can be embedded within an agent and used by that agent to help make decisions in particular situations.

The trust an agent has in another in a particular situation is related to the amount of trust in that agent in general and the importance of the situation to the trusting agent:

$$T_x(y, \alpha_x) = f(\, T_x(y) \,, I_x(\alpha_x))$$

In general, if $I_x(\alpha_x) > T_x(y)$, the resulting situational trust will be such that $T_x(y, \alpha_x) <$

$T_x(y)$. This is not always as binding as it may seem, however — note that the decision to trust a specific agent may also be related to the competence of that agent in the given situation, as observed or experienced in previous similar situations. In the examples below, the value of this function is obtained using the following equation:

$$T_x(y, \alpha_x) = T_x(y) + (T_x(y) * (T_x(y) - I_x(\alpha_x)))$$

In order to cooperate with agent y, the trust x has in y for that particular situation has to be above a certain threshold value, itself a function of the importance, costs and benefits of the situation concerned:

$$\text{If } T_x(y, \alpha_x) \geq \text{Cooperation\_Threshold}_x(\alpha_x) \Rightarrow \text{Will\_Cooperate}(x, y, \alpha_x)$$

where:

$$\text{Cooperation\_Threshold}_x(\alpha_x) = \frac{\text{Perceived\_Risk}_x(\alpha_x)}{\text{Perceived\_Competence}_x(y, \alpha_x)} \times I_x(\alpha_x)$$

Here, $\text{Perceived\_Competence}_x(y, \alpha_x)$ reflects what was discussed above, and allows for co-operation with an agent which is not trusted very much in general, or with an agent in an important situation, where that agent is known to be reliable and competent to a high standard in that or similar situations. The $\text{Perceived\_Risk}_x(\alpha_x)$ represents the agent's best estimate of the potential costs and benefits of the situation. Both of these are expanded as follows:

Perceived Competence

1. Since competence takes into account the agent to be trusted, the Perceived_Competence measure is based on experience in similar situations, experience of the same agent in similar situations, and knowledge of that agent's capabilities in similar situations. This will be illustrated further below. In this case, there is at present no formula devised for taking all of these into account. As such, in the examples to follow, it will be introduced as an arbitrary element.

2. In the second case, the trusting agent may know nothing of the other agent or the situation, in this case:

   $$\text{Perceived\_Competence}_x(y, \alpha_x) = T_x$$

   Note that it is possible for the trusting agent to know of the agent to be trusted. As such, it may have some form of trust in that agent already. In this case, instead of T_x, we could use the value of $T_x(y)$ as x's estimate of the competence of y in the situation. This is used in the first of the two examples below.

Perceived Risk

The Perceived_Risk for x is simpler. Risk involves a weighing up of the costs and benefits of situations — whether it is worth risking the costs in order to obtain the benefits of the situation being resolved. In the examples below, we use the following simple formula for estimating the risk involved in a situation:

$$\text{Perceived\_Risk}_x(\alpha_x) = \frac{C_x(\alpha_x)}{B_x(\alpha_x)} \times I_x(\alpha_x)$$

The cooperation threshold is higher the more the expected costs and the importance of the situation. Hence the more important the situation, the more trust is necessary to enter into a cooperative situation with another agent. If degrees of cooperation are possible, then this threshold could be *stepped* to take these degrees into account, and cooperation of a limited kind can be entered into with agents who are not trusted enough to cooperate with to the full extent required in the situation. Using the concept of trust alone is a limited approach to decisions such as whether or not to work with another agent. The idea of the importance of the situation takes this into account. Coupled with trust, this presents a powerful tool for an agent in decision making. Note that there may be situations where an agent has no choice but to trust another, for an example, see below. In this case, although the cooperation threshold may be high, cooperation will still occur. However, the agent can recognise the problems inherent in this, and make allowances, in the form of *safety nets*, in case of untrustworthy behaviour.

7.1 Examples of Cooperation

The formulae above are heuristic in nature, and, as was mentioned earlier, are not intended to represent the final workings of trust, in human or machine agents. They do, however, provide an agent with a useful tool for the evaluation of situations and potential situations and cooperative relationships. In order to illustrate this further, this section provides two examples of trust in action using the above formulae. The situations are adapted from Connah and Wavish, 1990 [6], and represent a problem involving furniture removal. In the first situation, only two agents are present, and there are two pieces of furniture to be moved. In the second, there are three agents and three pieces of furniture. Each agent in both situations has the goal of moving a piece of furniture (different from any other agent) to the door. Unfortunately, they cannot lift the pieces by themselves, and thus need help. A diagram of the start state for each situation is given below. For more information and an introduction to the problem, see [6].

7.1.1 Example 1: Two Agents

In this example (see Figure 1) we have two agents, each of which has the goal of moving a piece of furniture to the door.

Each piece of furniture is too heavy to lift for an agent alone. In other words, to achieve their goals, the agents must cooperate. This situation is therefore an example of agents not having any choice about cooperation. They must cooperate. Some interesting points, however, will hopefully arise. Table 2 shows the workings of the first agent (call it x) with regard to how much trust it has in the second agent (y). In this situation, x decides it can trust y and will cooperate in the situation. Interesting results can be obtained with different starting values for trust, costs, benefits, and situational importance. Consider, for example, a situation where the cooperation threshold is higher than the amount of trust x has in y. Unfortunately, although this should lead to x not cooperating at all, cooperation is necessary

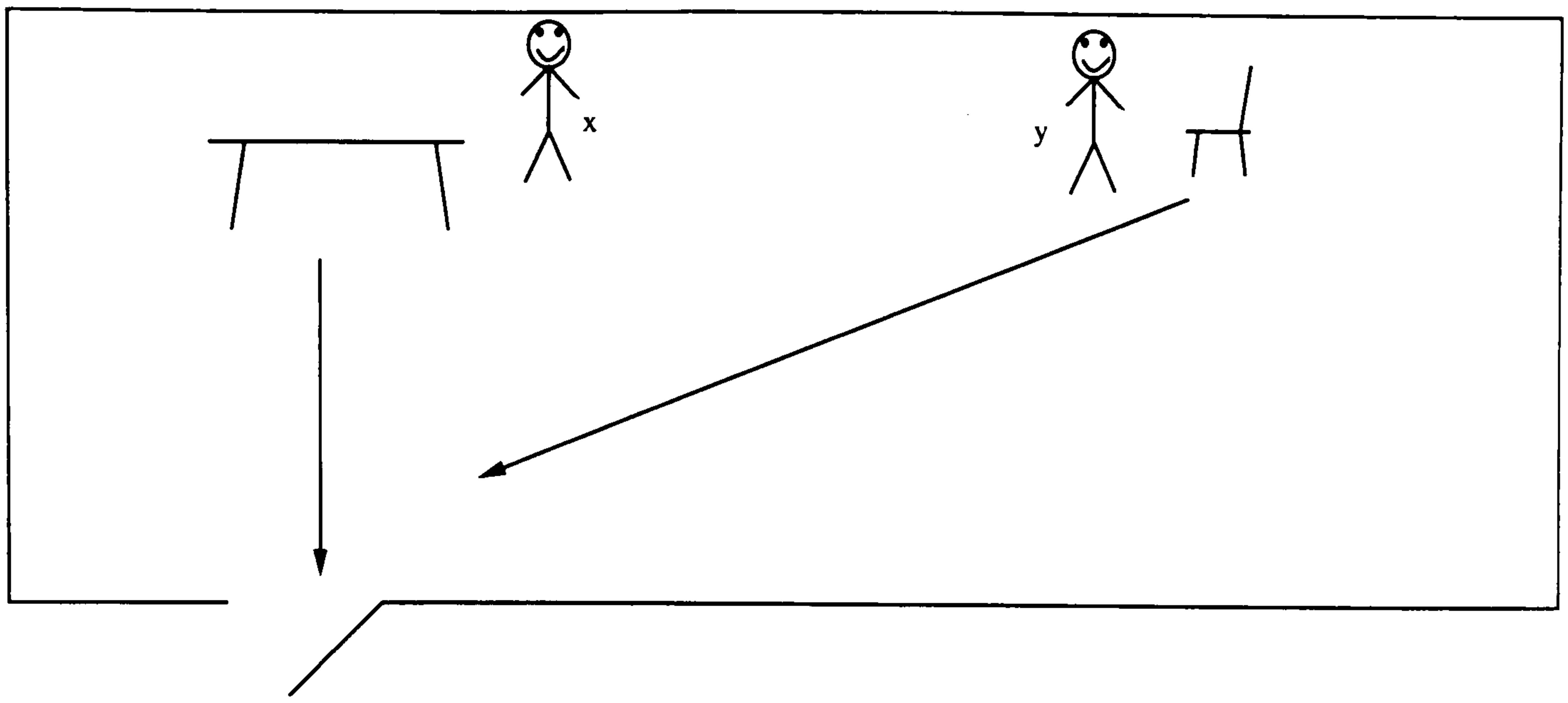

Figure 1: Furniture Moving — Two Agents at the Start of the Example

if the goal is to be fulfilled. Cooperation will therefore have to be entered into. The result of a high threshold, however, would allow x to monitor y's behaviour very closely, and protect itself, for example by insisting y help move x's furniture first, thus guaranteeing the result it wants.

7.1.2 Example 2: Three Agents

The above example illustrates how an agent can use trust to reason about relationships with little or no knowledge about the other agents concerned (consider if x didn't know y at all — she would just use the value of her general trusting disposition, T_x, in the place of $T_x(y)$). It also shows how some situations leave agents with little choice about cooperation — it is virtually coerced. In this case, perhaps the politics of coercion, rather than of trust, would be more applicable. The concept of trust becomes more interesting in a collection of agents, where different agents can be considered in the light of experience, competence, and so forth.

The present example considers the same situation, of furniture moving, but with three agents, and three pieces of furniture (see Figure 2).

We will consider the situation from x's point of view again. An interesting twist to the situation will be introduced in the fact that agent z is known and little trusted in general by x, but x 'knows' that z is, by chance, a specialised furniture removal agent, and thus extremely competent in this situation. Table 3 shows the 'thoughts' of agent x on cooperating with z.

The results of the deliberations are that x will favour working with y in this situation, despite z being a professional furniture mover. It is hoped that this mirrors real life situations to some extent — we may trust our best friend to help us move furniture rather than a professional, who may, perhaps, have less respect for our items of furniture.

8 Implementation

What has been presented thus far is a view of how trust can be represented in a fairly simple logical manner, but allowing for further expansion. It would be useful to discuss how a concrete implementation of such an idea could work. What is discussed in this section is at

Agent x trusts agent y 85%. She considers the situation to have an importance of 75% The Costs of the furniture not being moved amount, as far as x can estimate, to 60%, this is quite high, but the benefits of moving the furniture are estimated at 70% — higher than the costs of them not being moved.
$T_x(y) \ = \ 0.85$
$I_x(\alpha_x) \ = \ 0.75$
$C_x(\alpha_x) \ = \ 0.60, \ B_x(\alpha_x) \ = \ 0.70$

From the above values, x can work out the amount of situational trust in, or reliance on, y in this situation:

$$\begin{aligned} T_x(y, \alpha_x) &= T_x(y) + (T_x(y) * (T_x(y) - I_x(\alpha_x))) \\ &= 0.85 + (0.85 * (0.85 - 0.75)) = 0.85 + 0.085 \\ &= 0.935 \end{aligned}$$

This is in fact very high. Note from the above discussion that since initial trust in y is greater than importance of situation α_x, we would expect a high situational trust.

The risks associated with the situation are $\frac{C_x(\alpha_x)}{B_x(\alpha_x)} \times I_x(\alpha_x)$, and work out as follows:
$\text{Perceived\_Risk}_x(\alpha_x) \ = \ \frac{0.6}{0.70} \times 0.75 = 0.642$
These are relatively high, but the competence of the other agent in this situation is not known, neither is anything known about the situation from previous experiences.

Competence is thus taken as the trust x has in y:
$\text{Perceived\_Competence}_x(y, \alpha_x) \ = \ T_x(y) \ = \ 0.85$

From Risk and Competence, the threshold of cooperation can be found:
$\text{Cooperation\_Threshold}_x(y, \alpha_x) = \frac{0.642}{0.85} \times 0.75 = 0.567$
This is relatively low, since the trust x has in y in that situation works out as 0.935, a very high figure. As this is greater than the threshold, x decides that it is safe enough to cooperate with y in moving furniture.

Table 2: Agent x's 'thoughts' on trusting agent y

present unimplemented, and this is the next stage of the project. The notation presented above is intended to make it easier to envisage the a notion of trust within an agent, and in this section, some of the considerations involved in implementing the notion are briefly discussed.

Each new situation requires a re-evaluation, partial at least, of the amount of trust we have in the agent(s) with whom we are interacting. Clearly, the amount of trust we already have in the other agent(s) will play a large role in the judgements we make, indeed, should a similar situation have been carried through in the past with the same agent(s), our judgements are virtually made for us. Theoretically at least, it would be useful for our agents to have a knowledge of previous encounters with other agents *and* the situations, or context, of those interactions. In practice this may not be possible for reasons of space and speed, and so some other way of judging situations must be found. What should be remembered is a value relating to how we trust each different agent we have interacted with, which is alterable through our continuing experiences with these agents. The very least we can manage with is a general

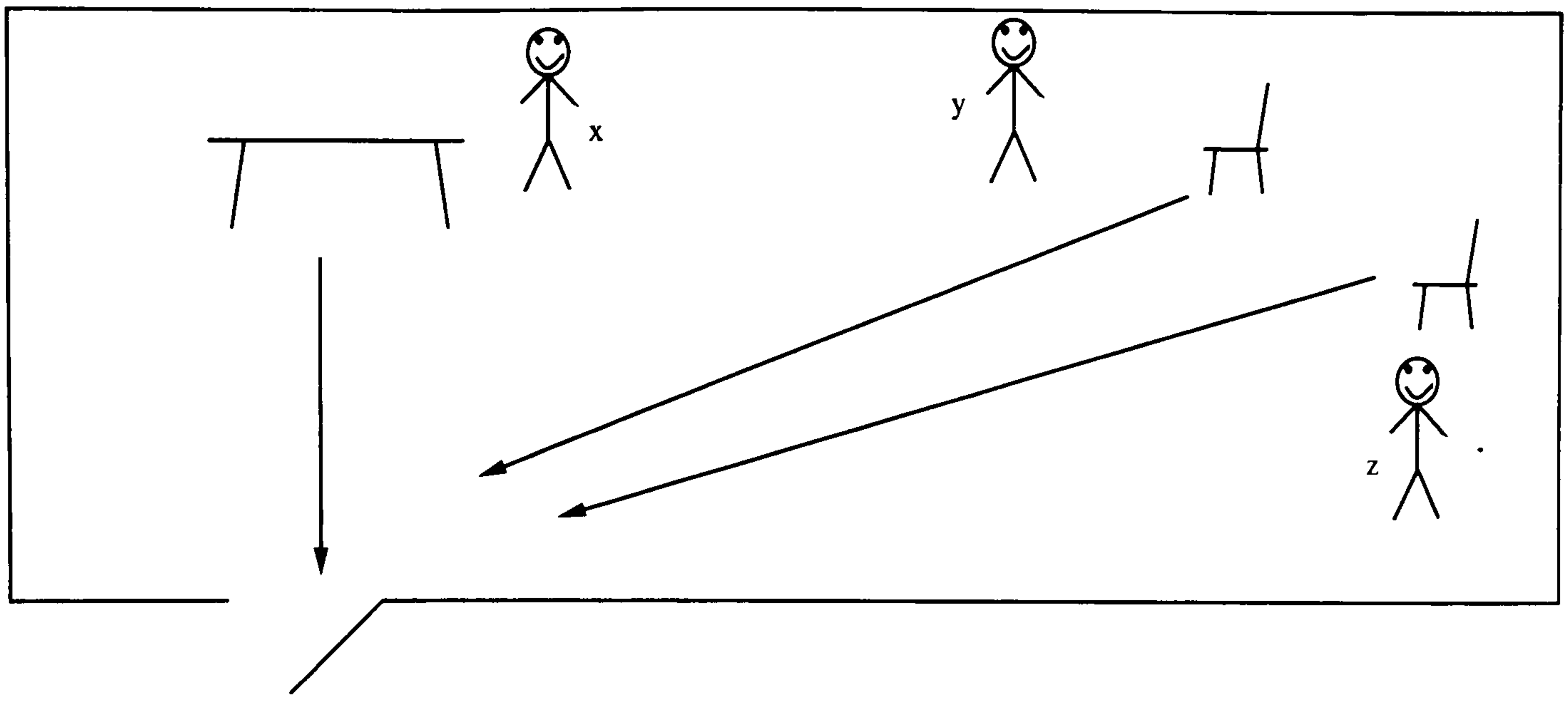

Figure 2: Furniture Moving — Three Agents at the Start of the Example

"trust" value which is altered with experiences with all agents, and applies to all agents.

In order to use a concept of another agent behaving 'well' or 'badly' in interactions, it is necessary to have some form of self-reference, such that the agent will know what it would have done in such a situation, and use this as a reference point for judging other agents. This self-reference would need to be embodied in our agents. An implementation of this could be based on finite state automata representations within the trusting agent. The trusting agent can attempt to model the behaviour of the trusted agent(s) using a simple finite state machine. The concept of behaving well or not then emerges simply from whether of not actual behaviour corresponds to the finite state machine represented by the trusting agent. This is an area for further consideration.

From the brief discussion above, it can be seen that the idea of trust is not task-specific, in other words it could be seen as an addition to what has been implemented thus far in any agent, and this late addition would not detract from either the workings of trust or the original task of the agent concerned.

A major task involved in the implementation is that of determining the workings of the various functions for altering the trust values discussed above. What has only been mentioned in passing thus far is one of the major aspects of trust: it is a *dynamic* reference to, in this example, interactions. The point is, it increases or decreases according to the behaviour of the trusted agents. For example, should y not help x after x had helped y move his piece of furniture, then the trust x has in y would decrease. Likewise, if y had helped, x's trust in y may increase. In addition, trustworthy or untrustworthy behaviour would alter the basic trust value of x (i.e. T_x) although not by as much (as was mentioned above, T_x is less dynamic than, for example, $T_x(y)$).

Whilst the concept of increasing or decreasing trust is clear, it is necessary to take care regarding the sensitivity of the trust values. Too sensitive, and the agents in question would never achieve a stable relationship; too static, and the agents concerned may well be not trusting enough or may be too easy to take advantage of. The ideal sensitivity will be arrived at through experimentation, and may well be task specific, contrary to the general idea of trust itself. In addition, the measure of sensitivity allows some control over the trusting behaviour of the agent.

Agent x trusts agent z 50%. The costs and benefits remain the same as in the above example. (Table 2)
$T_x(z) = 0.50$
$I_x(\alpha_x) = 0.75$
$C_x(\alpha_x) = 0.60, B_x(\alpha_x) = 0.70$

From the above values, x can work out the amount of situational trust in, or reliance on, z in this situation:

$$\begin{aligned} T_x(z,\alpha_x) &= T_x(z) + (T_x(z) * (T_x(z) - I_x(\alpha_x))) \\ &= 0.50 + (0.50 * (0.50 - 0.75)) = 0.50 - 0.175 \\ &= 0.325 \end{aligned}$$

This, compared to the situational trust in y of 0.935 (above), is very low. However, we must note that z is a furniture moving agent, and as such, although the risks involved in the situation stay the same, the competence will change dramatically.

The risks associated with the situation are as follows (see the main text for the formula):
$\text{Perceived\_Risk}_x(\alpha_x) = \frac{0.6}{0.70} \times 0.75 = 0.642$

Competence is taken as a fairly high figure, for this example. x decides z is 100% competent at moving furniture (z has the certificate to prove it!):
$\text{Perceived\_Competence}_x(y, \alpha_x) = 1.00$

From Risk and Competence, the threshold of cooperation can be found:
$\text{Cooperation\_Threshold}_x(y, \alpha_x) = \frac{0.642}{1.00} \times 0.75 = 0.482$
This is lower than that for agent y, which is 0.567. In other words, x would need to have less trust in z that in y to work with z in the furniture moving situation. In fact, this is the case, but the resulting trust x has in z is only 0.325, lower than the cooperation threshold for z in any case. x would therefore choose to work with y rather than with z.

Table 3: Agent x's 'thoughts' on trusting agent z.

9 The Limitations of Trust

The concept of trust introduced in this paper has its limitations. Not least is that the work is in its early stages, and that there is a considerable amount of work still to do. For example, thus far we have considered only one of the many different areas where trust can prove useful, namely that of initiating, or considering, cooperation with other agents. Whilst this is useful, it is only one area. Other important areas were briefly discussed above, notably the use of trust for gauging confidence in the information given to an agent, as in source control [10].

In addition, trust can help in the formation of groups: "Perhaps there is no single variable which so thoroughly influences interpersonal and group behaviour as does trust..." [12, page 131]. The formation of a group is basically an extended cooperative relationship between more than two agents. As such, the same kind of trust may well be used in forming the group. A question may arise in how groups themselves trust. For example, does a group trust another group more or less than another individual. Does a group see itself as a single entity, or just many individuals working to a common purpose? All these questions have yet to be addressed.

Trust provides an agent with a tool for judging the future, based on experience of the past. It is, however, limited when applied alone. It is not envisaged that an agent use trust as the sole measure of certainty in situations, much less as the only means of making a decision. The examples given above are intended to illustrate how trust can be used in such situations, but they are limited, as might be expected. When coupled with other methods, such as utility theory or theories of rational behaviour (e.g. [25]), trust provides an agent with a powerful and useful tool when interacting with other agents.

The heuristics given above present their own considerations. One of the problems with attaching values to things of that nature appears to be the different interpretations that can come from the values. It is hoped that the implementation of a trusting agent will enable a thorough investigation of these heuristics. They are, as such, introductory, but solid in that they work in a way in which trust would be expected to act in an agent, and allow for experimentation and discussion.

There are also problems with deadlock. Consider an agent in the furniture moving example who need help, but the only other agent around refuses to give it, and needs no help for itself (no furniture to move?). Nothing will happen — a classic deadlock situation, it seems. This suggests that such situations need reciprocity. This is, however, not so much a problem with trust as with cooperative situations in general. Nevertheless, it is a problem which will need addressing.

Finally, there is a large amount of complexity involved for one agent when considering many other in terms of trust. This may or may not be unavoidable, but it does suggest that trust may not be of much help in situated agents, where the behaviour is dictated by the environment. It will, however, help an agent when considering the environment and possible environment states, as briefly mentioned above.

10 Conclusions and Ideas for Further Work

The notion of trust presented here is extendable. Further work will concern the decisions an agent makes with regard to its environment; how the agent trusts what it interacts with in a physical sense. Secondly, the fact that an agent trusts another does not impel the trusted agent to behave in a 'proper', trustworthy manner, although the fact that the other agent is trusting implies an acceptance of this 'danger'. What would be useful is to include in the trusting world a law of some sort, not unlike the law-governed systems of Minsky [21, 20]. These systems provide laws which "...affect only what an agent may be able to do, not what it actually will do." [21, page 291]. What this means to the trusting agent is that the agent being trusted may be malevolent, but not capable of being malicious, and of actually damaging the trusting agent. In such a situation, trust may be built up more easily and safely, and this will be further studied in the future. In addition to the work discussed in section 9, further work will investigate the possibilities of using the concept of trust in, for example, distributed operating systems, as an analysis tool for more complex multi-system interactions, and so forth. Finally, the use of trust as a tool for the reduction of complexity, as briefly mentioned above, is an interesting avenue of work in itself.

In conclusion, a concept of trust has been introduced for agent interactions. The benefits of the concept include the allowance of a dynamic reference to interactions between agents, both with reference to the agent itself and to interactions with other agents. It allows at least minimal interactions and cooperation between agents where ordinarily there would be none. Most importantly the agents concerned are making an implicit acknowledgement of the possibility of malevolence or mistakes on behalf of the other agents, and as such there can be some form of backup against such occurences. It is thus of benefit both in that interactions are allowed, but with an acknowledgement of the problems involved. In the near future a working implementation of such a trusting agent will be forthcoming.

Acknowledgements

Many thanks to Ian Wilson, Harold Thimbleby, Lynne Coventry, Tom Kane and others for constructive comments on earlier drafts of this paper. Without whom... The comments from anonymous referees enabled the author to improve this paper a great deal, for which, many thanks. The author is a SERC CASE student, partially supported by Canon Research Europe.

References

[1] Robert Axelrod. *The Evolution of Cooperation.* Basic Books, New York, 1984.

[2] Robert Axelrod. The evolution of strategies in the prisoner's dilemma. In Lawrence Davis, editor, *Genetic Algorithms and Simulated Annealing*, pages 32–41. Pitman, London, 1987.

[3] Roy L. Behr. Nice guys finish last - sometimes. *Journal of Conflict Resolution*, 25(2):289–300, June 1981.

[4] Richard Boyle and Phillip Bonacich. The development of trust and mistrust in mixed-motive games. *Sociometry*, 33:123–139, 1970.

[5] Stephanie Cammarata, David McArthur, and Randall Steeb. Strategies of cooperation in distributed problem solving. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1983.

[6] David Connah and Peter Wavish. An experiment in cooperation. In Yves Demazeau and Jean-Pierre Muller, editors, *Decentralized AI*, pages 197–212. Elsevier Science Publishers (North-Holland), 1990.

[7] Peter A. Danielson. Artificial Morality: Prolog and the Prisoner's Dilemma. Corrected version of a paper presented at the Fifth International Conference on Computers and Philosophy, Stanford University, 8–11th August, 1990., 1990.

[8] Morton Deutsch. Cooperation and trust: Some theoretical notes. In M. R. Jones, editor, *Nebraska Symposium on Motivation.* Nebraska University Press, 1962.

[9] Julia Rose Galliers. A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict. Technical Report No. 172, University of Cambridge Computer Laboratory, 1989.

[10] Roberto Garigliano, Albert Bokma, and Derek Long. A model for learning by source control. In Bouchon, Saiger, and Yager, editors, *Uncertainty and Intelligent Systems*, pages 163–170. Springer Verlag, Lecture Notes in Computer Science, LNC 313, 1988.

[11] Matthew L. Ginsberg. Decision procedures. In M. Huhns, editor, *Distributed Artificial Intelligence, Volume 1*, chapter 1, pages 3–28. Pitman, London, 1987.

[12] Robert T. Golembiewski and Mark McConkie. The centrality of interpersonal trust in group processes. In Cary L. Cooper, editor, *Theories of Group Processes*, chapter 7, pages 131–185. Wiley, 1975.

[13] Lars Hertzberg. On the attitude of trust. *Inquiry*, 31(3):307–322, September 1988.

[14] Nick R. Jennings. On being responsible. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany, Panel Session*, 1991.

[15] John T. Kohl. The evolution of the Kerberos authentication service. In *Europen '91 - European Conference on Open Systems*, pages 295–313, 1991.

[16] Olli Lagenspetz. Legitimacy and trust. *Philosophical Investigations*, 15(1):1–21, January 1992.

[17] Bernhardt Lieberman. *i*-Trust: a notion of trust in three-person games and international affairs. *Journal of Conflict Resolution*, 8(3):271–280, 1964.

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

[19] Steven McNeel. Training cooperation in the prisoner's dilemma. *Journal of Experimental Social Psychology*, 9:335–348, 1973.

[20] Naftaly H. Minsky. The imposition of protocols over open distributed systems. *IEEE Trans. Software Engineering*, pages 183–195, February 1991.

[21] Naftaly H. Minsky. Law-governed systems. *Software Engineering Journal*, pages 285–302, September 1991.

[22] Bonnie M. Muir. Trust between humans and machines, and the design of decision systems. *International Journal of Man-Machine Studies*, 27(5 & 6):527–539, Nov/Dec 1987.

[23] Chisato Numaoka. Conversation for organisational models. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany, Panel Session*, 1991.

[24] Derek Parfit. Prudence, morality, and the prisoner's dilemma. In Jon Elster, editor, *Rational Choice*, pages 34–59, 1986.

[25] Jeffrey S. Rosenschein. *Rational Interaction: Cooperation among Intelligent Agents.* PhD thesis, Stanford University, 1985.

[26] John F. Shoch and Jon A. Hupp. The "worm" programs: early experience with a distributed computation. *Communications of the ACM*, 25(3):172–180, March 1982.

[27] Eugene H. Spafford. The Internet Worm: Crisis and Aftermath. *Communications of the ACM*, 32(6):678–687, June 1989.

[28] Robert L. Swinth. The establishment of the trust relationship. *Journal of Conflict Resolution*, 11(3):335–344, 1967.

[29] Harold Thimbleby. Can viruses ever be useful? *Computers and Security*, 10:111–114, 1991.

[30] Donnell Wallace and Paul Rothaus. Communication, Group Loyalty, and Trust in the PD Game. *Journal of Conflict Resolution*, 13(3):370–380, 1969.

[31] Thomas Y. C. Woo and Simon S. Lam. Authentication for distributed systems. *IEEE Computer*, pages 39–52, January 1992.

Trust in DAI — A Discussion

Stephen Marsh
Department of Computing Science and Mathematics
University of Stirling
Stirling
FK9 4LA
SCOTLAND
email: spm@cs.stir.ac.uk
Telephone: (+44) 786 67444

April 22, 1993

Abstract

A discussion of trust is presented which focuses on multi-agent systems, from the point of view of one agent in a system. The roles trust plays in various forms of interaction are considered, with the view that trust allows interactions between agents where there may have been no interaction possible before trust. Trust allows parties to acknowledge that, whilst there is a risk in relationships with potentially malevolent agents, some form of interaction may produce benefits, where no interaction at all may not. In addition, accepting the risk allows the trusting agent to prepare itself for possibly irresponsible or untrustworthy behaviour, thus minimizing the potential damage caused. A formalism is proposed to allow agents to reason with and about trust[1].

Contents

[1] This is a revision of a paper presented at the 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Italy, 1992 [20].

1 Introduction

Distributed Artificial Intelligence (DAI) is concerned primarily with cooperation: why it happens, who cooperates with whom, when, and to what extent. However, DAI has thus far concentrated on why cooperation takes place, why it is useful to an agent to cooperate, with little thought given to any of the other aspects of cooperation. In particular, little has been done with a view to how agents will survive in the 'outside world', as opposed to the restricted experimental worlds they exist in today. Cooperation is generally accepted as being a good thing [2, 5, 7]; however, whether or not this is the case, there will be situations where agents have conflicts of interest, and some form of coping with these conflicts is required [28, 29]. Whilst some studies take this into account, [6, 29, 28], they make the important assumption that the agents are trustworthy. This is indeed seen as "absolutely essential" [29], at least in a situation where the agents are communicating, and may well be more important if communication is not allowed. The agents that we design for use within our experimental worlds will most likely not be robust enough to survive outside the laboratory. Why is this? Altruism, however desirable, is not the "name of the game" in the world we live in. Indeed, taking the world of computing alone, malevolence, not altruism, appears to be prevalent [33], and this is irrespective of how well-intentioned the work may be to start with ([30] gives details of the idea of a useful "worm" program which was used to devastating effect not so long ago [31], despite never being intentionally designed for malevolent purposes).

Taking this lack of altruism into account, some measures must be taken to make our agents less vulnerable to others' incompetent or malevolent behaviour. There are different approaches to this. The first, most obvious, is not to interact with anyone we don't know, and it is this method which is used in the simple username, password schemes on computers, and other security systems such as Kerberos [16]. This approach is too restrictive to allow a normal interaction between agents. Another method is to ignore the vulnerability of our agents, and hope they will not be exploited. This, too, is shortsighted. I propose that there is a continuum between these two extremes which can be used to the advantage of agents, and that the points on this continuum represent varying degrees of trust on behalf of an agent.

2 Why Trust?

The concept of trust may seem a little unusual to suggest for computers; it is thus worthwhile to put forward some reasons why it may prove useful. Implicit in the notion of Distributed Artificial Intelligence is the concept of decentralisation. Since decentralisation implies a lack of central control, and with it a lack of guidance in the 'right' direction, it becomes necessary, in order to carry through successful interactions with other agents, to develop some judgement as to the worth of these interactions and the risk associated with them. The concept of trust has already been widely field-tested with respect to the human race [9, 4, 32, 34]. It provides us with an ideal measure of expectation of risk, and since the risk is implicitly acknowledged in the form of a trusting relationship then some form of measure can be taken against untrustworthy actions, which might otherwise be fairly damaging — a form of 'safety net'. In a laboratory, we, and our agents, are acting under controlled conditions with the knowledge of what kind of behaviour to expect, and this knowledge comes forward and is instilled, unconsciously or not, in our experiments. In the case of DAI, the agents that are built are as subject to this 'rule' as anything else: "it makes little sense to ask *why* they are helping one another; they help each other because they have been designed that way" [29, page 12]. What the concept of trust can help us ensure is that our agents are more robust with respect to interactions with agents that are not our own, and interactions of a type that is not forseen. In addition, the concept of trust in certain future eventualities and in the expected behaviour of the environment (including other agents) functions as a tool for the reduction of complexity for the trusting agent [19]. This will be discussed further below.

Assuming blindly that cooperation is a good thing is not necessarily the correct approach, although this depends on the viewpoint of our agents. Behr [3] points out that even something as conceptually simple as the Prisoner's Dilemma [27, 22, 2, 1] can be viewed from differing perspectives, depending on whether one wishes to score highly, as Axelrod had assumed [2], or whether one wishes to defeat the opponent. In the latter case, the more successful strategies were not so nice, since to win, one must "defect more than one's opponent does" [3]. The moral of this exercise, if there is one, is that if we assume that everyone is out to cooperate, we are mistaken. Lack of cooperation may actually benefit others. What is perhaps better to assume is a notion of self-interested agents, all out to get the best they can [18]. In some cases that means that they cooperate, in others, their behaviour can range from non-cooperation to downright maliciousness. Since we cannot presume to know what their behaviour may be at any given time, the notion of trust put forward in this paper relies on a judgement based on experience, coupled with, if available, past knowledge of the agent to be trusted and their behaviour.

3 What it is

Trust implies a risk of some sort, "One trusts when one has much to lose and little to gain" [9, page 304]. From many of the definitions, this is taken to be the case, as in [32], who states that entering a trusting relationship is "choosing to take an ambiguous path that can lead to a beneficial event or a harmful event depending on the behaviour of the other person — where the harmful event is more punishing than the beneficial event is rewarding."

Trust is an emotive issue. Its definitions also are based mostly on emotion, or on moral responsibilities toward the trusting party [25]. It is worth reminding ourselves that computers as such cannot feel the moral responsibilities that humans do, although interesting work is being done in this area [8]. If this is the case, a more general definition of the concept of trust is required, one which does not rely entirely on moral bounds and expectations, but also on an expectation on rationality on the part of trusted agents. Lieberman introduced such a concept in 1964 [18], which he used to apply to inter-nation relationships. Called *ı*-trust, it is based on a theory that nations will act in a self-interested way (hence the *ı*), and this could be relied upon to trust them. Interestingly enough, seemingly irrational behaviour can in fact be predictable in that immediate small gain may be passed over for later large gain, and it is this that is the central thesis of *ı*-trust.

Lieberman's definition of trust is "a belief or expectation about behaviour in a situation in which the problem of forming a stable coalition structure is important [...] the belief that the parties involved in the agreement will actually do what they have agreed to do; they will fulfill their commitments not only when it is obviously advantageous to do so, but even when it may be disadvantageous, when they must sacrifice some immediate gain." This is because the interests of an agent "transcend the increased immediate gain he might make if he defected from a coalition [...] He keeps agreement so that he will be trusted, so that his partner in turn will stay with him and the coalition will grow rich"[18, page 279].

3.1 Security

Between blind trust and complete mistrust lies a continuum of varying degrees of trust. The security measures of today offer a mixture of both blind trust and complete mistrust. As an example, if I am a complete stranger attempting to access a computer, if I didn't know a password, I will not be allowed access. If, however, I stumble across a password, or obtain another method of entry, then the computer affords me almost complete trust, limited only by the prior actions of the system administrator. The same applies to agents — they may be initially completely non-trusting, but once the outer defences are penetrated, they can be used in any way, to transmit viruses, for example.

The concept of dynamic trust enables an agent or system to interact (at least to some extent) with some other agent(s). The limits of interaction change with experience of the action of the other agent(s). In other words, an agent can interact with others to a certain extent, trusting them, or relying on them, to that extent. In the light of trustworthy behaviour on their part, the extent to which the agent trusts them (and hence to which it is prepared to interact with them) will increase, and untrustworthy behaviour results in a decrease in the amount of trust.

Why would we want to trust a stranger? The answer lies in what we are expecting. From any relationship, there is some way of benefiting for both, or all, of the members of the relationship. In information-sharing, for example, all information may well be useful, and finding out which is useful and which is not depends on having the information to hand. If I were not to trust ninety percent of people, I may well lose what I would consider vital information. This will be discussed further in the following section, along with some other situations where trust allows for a useful and potentially beneficial interaction to take place in a situation where no trust at all would leave our agents lacking in some way.

4 Reliance — Kinds of Trust

Different situations require different forms of action. While I may trust you to drive me to an airport, trusting you to fly the plane is another matter! What this implies is that for different interactions between agents, a different kind, or form, of trust may be required, in that different things need to be taken into account with regard to different situations. Note that this is not the same as having different *degrees* of trust in an agent, as discussed above. Indeed, there can be different degrees of the different *kinds* of trust. This is closely coupled to the agents view of a situation, as presented below.

The amount of trust in a person or agent does not change from situation to situation solely because the situation changes; rather, the reliance, based on the trust in the agent, changes according to the different situation. Hence, if I were to speak in terms of percentages of trust or reliance, I may say I trust (rely on) you 50% to drive me to the airport, but only 20% to fly the plane. The amount of reliance I may have in you at a given moment, or in a given situation, is a function of the amount of trust I have in you in general, in addition to my experience of your actions in similar situations in the past, and the competence I perceive you to have in the situation. In some situations, however, the *value* we place on trust is higher than in others. As an example, if I were in a situation where a wrong decision could cost me my life, I would think very hard about taking your advice, especially if I didn't trust you completely.

The following subsections suggest some examples of different situations in which reliance is involved.

4.1 Cooperation

The advantages of cooperation are many and diverse. Indeed, many tasks cannot be performed by one agent alone, perhaps because the agent does not have all of the knowledge necessary to formulate a solution to the problem, perhaps because the agent cannot physically perform a task without help [7], and so forth. In its simplest form, a cooperative relationship is between two agents. In order to initiate a cooperative relationship, some form of trust is required, "the initiation of cooperation requires trust whenever the individual, by his choice to cooperate, places his fate partly in the hands of others."[9, page 302]. The amount of trust, and the importance of it, depends on the situation the relationship is formed to handle.

4.2 Information Sharing and Gathering

Consider an agent seeking information in the world at large. As such, this agent may well be given information by other, unknown agents. This information has to be judged in some way, and the notion of trust may play a part here. In other words, the agent can use a measure of trust in the agent giving the information, and also the validity of the information concerned. This latter measure could be based on the prior knowledge of the agent, and the way in which the information correlates with what is already known. In this case, a knowledge of or trust in the other agent may not be necessary; for example, if I were to show you a white piece of paper and tell you it was white, all you would need to trust was your own judgement. Some information may also be this self-evident. Using already known and accepted knowledge may not be possible, since nothing may be known about the information in question, and thus there is a useful 'backup' with the measure of trust in the agent giving the information. For example, were I to put a piece of paper behind a chair where you could not see it, and tell you it was white, you would have to trust me in order to believe this information in that situation. Trust would allow you to accept this with some measure of certainty or uncertainty. Indeed, the work of Garigliano *et al* in the area of Source Control is an example of this to some extent [10].

4.3 Trust as Reduction of Complexity

As was mentioned briefly above, the concept of trust acts as a tool for the reduction of complexity for an agent as regards future eventualities. That is to say, an agent need only consider trusted eventualities, or world states that arise from trusted actions and world states. In other words, we do not have to consider all possibilities if we trust that some will not happen. That does not, however, mean that we will not prepare for them. We can still take precautions, even against something we know nothing of. The results of this is that the amount of processing devoted to considering future events is likely to be drastically reduced. This would be a great help for an agent as it makes a plan. At present, the idea is very much in it's early stages. It is hoped, however, that an agent architecture exploiting trust as a tool for the reduction of social and environmental complexity can be implemented.

5 Relation to Other Work

Numaoka [26], presents a view of the Special Interest Group, or SIG, which he states has "the most fundamental style of every organisation." A basic proposal in his paper is the idea of agents within a group voting to let another agent join the group. This could be extended using our definition of trust proposed here, such that the vote can be influenced, if not determined, by the trust the members of the group have in the prospective new member. This, in fact, provides an insight to the theory proposed here, in terms of group trust, which could be seen as the amount of trust a group has in an individual, whether inside or outside of the group, and also in another group (again, inside or outside of the group in question). Indeed, social choice is an aspect of decision making that may well benefit from a formal definition of trust.

Jennings [15] provides us with a view of "Joint Responsibility", in terms of what an agent states it will do, and the responsibility it has to the group as a whole. This, if developed further, could be used to influence the trust factor of our agents. For example, if an agent does not carry out the tasks it says it will, and is thus irresponsible, then we can use this irresponsibility as some form of metric towards altering our trust values accordingly.

There is some concern over the problem of security in distributed systems, in that "...the notion of trust in distributed systems is poorly understood. A satisfactory formal explication of trust has yet to be proposed."[35, page 40]. The idea of trust provided in this paper may go some way towards helping in that area.

6 A Formalism

Thus far in the report, we have proposed areas of work where a consideration of trust is of use. The remainder of the paper will introduce a formalism for trust which will allow a DAI agent to reason with and about trust. The justifications for such a formalism are, at the least, threefold. First, a formalism can help to avoid ambiguities in discussing a concept. For a concept as value-laden as is trust, the formalism can help avoid many possible disagreements by making clear what is actually being discussed. Second, if we are to progress to actually implementing some kind of trust apparatus within an agent, the formalism can help in that implementation, since it may be a small step from there to a working example. Third, the possibility of actually having worked examples can help us to determine whether or not the proposed heuristics actually work. For example whether incrementing trust by such and such an amount following trustworthy behaviour is justified, or if the method of using trust in decisions is satisfactory. With a formalism and some heuristic formulae, it is possible to determine experimentally and theoretically the answers to these and other questions. There may be other benefits to a formalism — the discovery of these will be left to the reader, since it is hoped that the ideas presented in this paper will be taken up for discussion and thought.

Before introducing the notation, I provide a disclaimer. As suggested above, trust is a value-laden concept, which everyone knows of and uses, most of the time [19]. The notation presented here is thus not intended to be a definitive view of the way trust and reliance work. It provides a useful discussion article and is intended to initiate interesting, productive debates on the topic. If the discussion and the notation are too weak for some, they have an invitation to expand or contract it, and to let me know of their thoughts and comments. In addition, the formalism proposed in this report is in no way final: it is an experimental piece, and continuously in flux. As presented below, it applies to a cooperative situation between two agents from the point of view of one of the agents. For situations involving more that two agents, or other aspects of cooperation, the formulae may well need adjustment. The intention is to prove that such considerations are indeed possible with trust, and that it provides a helpful addition to the decisive powers of agents.

7 Notational Definitions

7.1 The Agents

We represent particular agents by the letters a to z. Each agent is a member of $\mathcal{A}$, the set of all agents. An agent can be considered to be an independent entity extant in a world populated with other such entities, each of more or less complexity than itself. In the introductory notation presented below, we present each formula unrelated to any other.

7.2 Situations Agents Find Themselves In

A situation is defined as a point in time relative to a specific agent. This allows us to consider two or more agents, at the same point in time, and even in the same place, to be in different situations, since each agent may have different knowledge, beliefs, or intentions at that time, colouring their perception of the situation. For example, consider the example given in [13], where there are several children playing, some of whom have mud on their foreheads. For each child, the situation is different, since they do not know if they have mud on their forehead (they may have), but they can see all the other children. In a situation where knowledge may be partial, that situation is seen differently by all in it. So in our notation situations are represented as the greek letters, α to ω, with a suffix representing the agent concerned (the agent who sees that situation from her point of view), so we have α_x to ω_x for situations from the point of view of agent x.

7.3 A Basic Trust

An agent, as a trusting entity, has a basic trust 'value', derived from previous experience. This value, which is dynamically altered in the light of all experience, is used a great deal in the formation of trusting relationships with unknown agents, and is represented as T_x, and has a range over $(-1, 1)$. A value of -1 represents total distrust, 0 is not trusting at all[2], and a value of 1 indicates total trust[3]. It is important to realise that it does not correspond to the amount of trust x has in any specific agent, but only to the general trusting 'disposition' of x. Since it represents this basic trust, it is less 'fluid' and changeable than a trust in any specific agent.

7.4 The Trust Value

Given two agents, $x, y \in \mathcal{A}$, to say x trusts y, we write: $T_{x.y}$. In addition to being a representation of the fact that x trusts y, this is a value, over $(-1, 1)$, of the amount of trust x has in y[4]. In other words, should $T_{x.y} = 1$, then x has complete trust in y. The trust value is a view of a particular agent of another with regard to the trusted agents general capabilities. As discussed briefly above, different situations may require different views of trust. The amount of trust in a particular agent in a given situation is represented here as, for example, $T_{x.y}^{\alpha_x}$, over (-1,1) for x's *situational* trust in, or reliance on, y to perform correctly in situation α_x. Note the interchangeability of the concepts of situational trust and reliance here. At present, they are considered to be identical. However, it may be feasible in the future to separate the two concepts, although closely related, to reflect more closely the more philosophical views on the subject [17, 14].

7.5 Further Notation

The agent's estimate of how important a situation is to itself is a value over (0,1), represented by $I_x^{\alpha_x}$ here. The importance value in this example is x's estimate, or subjective measure, of how much situation α_x means to it. A value of 0 means that the situation is of no importance, and as such there are no risks associated with it, whereas a value of 1 represents a situation of the utmost importance. The importance of a situation to an agent is useful in determining the amount of situational trust to place in an agent at any given time, as will be seen below; as an example, importance could be measured in terms of payoff functions [29, 11]. In the examples below, it will be introduced as an arbitrary value. Further models of trust will expand the idea of importance, and how it can be estimated by the agent.

[2] Note that these are not the same. Not trusting someone is not the same as having no opinion on the matter. Put another way, distrust $x \not\equiv$ not trust x.

[3] I prefer not to call this 'blind' trust, since that implies a lack of consideration, or careful thought, on the part of the truster. In the system described here, this is not the case.

[4] It is realised that attaching values to things creates its own problems, such as "what is the difference between a trust of 0.51 and a trust of 0.52?" This is acknowledged, but the values are at present retained for two reasons: Firstly, attaching a value to something allows us to talk in quantitative and qualitative terms, such as, "a trust of 0.75 is quite high, and higher than a trust of 0.70". Secondly, the use of values allows us to use calculations in arithmetic such as those introduced below. The difficulties remain, but it is hoped that the benefits outweigh these.

Related to the importance of a situation are the concepts of costs and benefits pertaining to that situation. The costs of a situation are measured in terms of the problems associated with incompetent or malevolent behaviour on the part of another agent in a relationship. Initially, the agent can only estimate the potential costs of a situation, based on past experience of similar situations, or just a 'rule of thumb' if that situation is new to the agent. They are represented here by $C_x^{\alpha_x}$, with a value over (0,1). Note that any agent(s) involved in the situation are not represented. This is because the potential costs of a situation are relative only to the agent concerned. As such, whoever is to be trusted and cooperated with, the potential costs of untrustworthy behaviour remain the same.

The benefits of a situation, or at least the expected benefits of trustworthy behaviour from the agent(s) being worked with, play a large part in decisions of whether or not to cooperate in the first place. In the notation here, the benefits are represented as $B_x^{\alpha_x}$, over (0,1).

As was mentioned above, the costs and benefits of a situation are related to the importance of that situation to any agent. Importance goes further than a simple weighing up of costs and benefits, however, and may include some knowledge or assumption about future benefits, preparation for further cooperation, and so forth.

Since trust is based on an agent's experience of previous interactions and situations to a large extent, and subjective in that it depends on individuals, some method of showing whether an agent is known to another or not is needed. This is referred to as *acquaintance* in that an agent becomes acquainted with another. Although there are degrees of acquaintance, for simplicity the examples below will treat it as a boolean concept — an agent either knows another or not. The concept of knowledge, or acquaintance, is represented as $K_{x \cdot y}$, which signifies the fact that x knows agent y if it equals 1, and not otherwise.

A summary of the notation used in this paper is presented in table 1.

Situations are represented by $\alpha_x \ldots \omega_x$.
Individual agents are represented by $a \ldots z$, and are members of $\mathcal{A}$, the set of all agents.
Basic Trust Value for Agent x: T_x
General Trust x has in y: $T_{x \cdot y}$
Situational Trust (Reliance) x has in y in situation α_x: $T_{x \cdot y}^{\alpha_x}$
Importance of situation α_x to agent x: $I_x^{\alpha_x}$
Potential costs to agent x following untrustworthy behaviour from another trusted agent in situation α_x: $C_x^{\alpha_x}$
Potential benefits to agent x following trustworthy behaviour from another trusted agent in situation α_x: $B_x^{\alpha_x}$
Representation of whether agent x knows (is acquainted with) agent y: $K_{x \cdot y}$

Table 1: Summary of notation for Trust — See the text for more details.

8 Rules for Interactions

Using the notation introduced above we can reason about trusting relationships. An example is given below of how trust can be used to determine the value of a cooperative relationship to a trusting agent. The examples presented below are intended to demonstrate how a particular agent can reason about the future using trust. The equations and the values associated with them are not intended to represent the cognitive working of trust in humans, and neither should they be taken as a final statement of how trust works within agents. They are intended to illustrate the thesis that a theory of trust, and approximate workings of that theory, can be embedded within an agent and used by that agent to help make decisions in particular situations.

8.1 Situational Trust

The trust an agent has in another in a specific situation is a function of the amount of trust in that agent in general and the importance of the situation to the trusting agent. In addition, the trust the other agent has in the first may play a part — knowing that you trust me may help me to reciprocate that trust, as does an estimate of how much I think you may trust me[5]:

$$T^{\alpha_x}_{x.y} = f(T_{x.y}\,,\, I^{\alpha_x}_x\,,\, \mathrm{Est}_x(T_{y.x}), \mathrm{Est}_x(\mathrm{Est}_y(T_{x.y})))$$

Of course, x can only estimate the amount of trust y has in her ($\mathrm{Est}_x(T_{y.x})$). This estimate may be wildly wrong, but could be quite sophisticated, taking into account what other agents may have said, previous behaviour of y, and so forth. Likewise, the estimate of expected trust ($\mathrm{Est}_x(\mathrm{Est}_y(T_{x.y}))$) is useful. The recursion here may cause problems of infinite regress. However, the extra information attained from each level of recursion becomes of less and less significance. It has thus been decided to limit the level of recursion.

In general, the agent's conception of the importance of the situation plays a larger part in determining the amount of situational trust than does the estimates described above. In the following formula, a high situational importance (positive) tends to decrease the situational trust with regard to general trust, and a low (negative) importance tends to increase it, since the situation is not important enough to worry. This is not always as binding as it may seem, however — note that the decision to trust a specific agent may also be related to the competence of that agent in the given situation, as observed or experienced in previous similar situations. There are also other ways of looking at such things — if the importance of a situation is high, then it may be the case that the agent has no choice but to cooperate. This does not, however, affect how much trust is placed in the other agent, and may help the trusting agent to recognise the need for some form of 'safety net' (see below).

We use the following formula to calculate the value of situational trust:

$$T^{\alpha_x}_{x.y} = \left(T_{x.y} - \frac{I^{\alpha_x}_x}{A}\right) + \left(\frac{\mathrm{Est}_x(T_{y.x})}{B} + \frac{\mathrm{Est}_x(\mathrm{Est}_y(T_{x.y}))}{C}\right) \tag{1}$$

The values of A, B, and C are left unfilled here, but it is suggested that the working of the equation is better if the inequality $A \geq B \geq C$ holds. The system described here uses the values of $A = 2$, $B = 4$, and $C = 5$. This reflects the idea that, when considering situational trust, general trust is more important than the actual importance of a situation, and also that, since estimates are at times less than accurate, their importance in working out situational trust should be consequently reduced by some factor.

8.2 Cooperation Threshold

In order to cooperate with agent y, the trust x has in y for that particular situation has to be above a certain threshold value, itself a function of the importance, costs and benefits of the situation concerned:

$$\text{If } T^{\alpha_x}_{x.y} \geq \text{Cooperation\_Threshold}_x(\alpha_x) \Rightarrow \text{Will\_Cooperate}_x(y, \alpha_x)$$

[5] And so forth. This recursive nature of trust may be problematic. The final formula takes into account just one level of recursion (see equation 1). This is, of course, expandable, but a limit point may exist somewhere.

where we take:

$$\text{Cooperation\_Threshold}_x(\alpha_x) = \frac{\text{Perceived\_Risk}_x(\alpha_x)}{\text{Perceived\_Competence}_x(y, \alpha_x)} \times I_x^{\alpha_x} \quad (2)$$

Here, $\text{Perceived\_Competence}_x(y, \alpha_x)$ reflects what was discussed above, and allows for cooperation with an agent which is not trusted very much in general, or with an agent in an important situation, where that agent is known to be reliable and competent to a high standard in that or similar situations. The $\text{Perceived\_Risk}_x(\alpha_x)$ represents the agent's best estimate of the potential costs and benefits of the situation. Both of these are expanded as follows:

8.2.1 Perceived_Competence

1. Since competence takes into account the agent to be trusted, the Perceived_Competence measure is based on experience in similar situations, experience of the same agent in similar situations, and knowledge of that agent's capabilities in similar situations. These considerations are naturally quite subjective. It is possible to visualise an agent keeping some form of database which records situation types, agents within those situations, and the observed competence of those agents in those situations. At present, such an implementation is incomplete. As such, the following, second case of Perceived_Competence is used in this paper.

2. In the second case, the trusting agent may know nothing of the other agent or the situation, in this case:
$$\text{Perceived\_Competence}_x(y, \alpha_x) = T_x \quad (3)$$
Note that it is possible for the trusting agent to know of the agent to be trusted, but not within the particular situation they find themselves in. It follows that the agent may have some form of trust in the other already. In this case, instead of T_x, we could use the value of $T_x\,y$ as x's estimate of the competence of y in the situation. i.e.:
$$\text{Perceived\_Competence}_x(y, \alpha_x) = T_x\,y \quad (4)$$
This is used in the first of the two examples below.

8.2.2 Perceived Risk

The Perceived_Risk for x is simpler. Risk involves a weighing up of the costs and benefits of situations — whether it is worth risking the costs in order to obtain the benefits of the situation being resolved. In the examples below, we use the following simple formula for estimating the risk involved in a situation:

$$\text{Perceived\_Risk}_x(\alpha_x) = \frac{C_x^{\alpha_x}}{B_x^{\alpha_x}} \quad (5)$$

The cooperation threshold is higher the more the expected costs and the importance of the situation. Hence the more important the situation, the more trust is necessary to enter into a cooperative situation with another agent. If degrees of cooperation are possible, then this threshold could be *stepped* to take these degrees into account, and cooperation of a limited kind can be entered into with agents who are not trusted enough to cooperate with to the full extent required in the situation. Using the concept of trust alone is a limited approach to decisions such as whether or not to work with another agent. The idea of the importance of the situation takes this into account. Coupled with trust, this presents a powerful tool for an agent in decision making. Note that there may be situations where an agent has no choice but to trust another, for an example, see below. In this case, although the cooperation threshold may be high, cooperation will still occur. However, the agent can recognise the problems inherent in this, and make allowances, in the form of *safety nets*, in case of untrustworthy behaviour.

8.3 Examples of Cooperation

The formulae above are heuristic in nature, and, as was mentioned earlier, are not intended to represent the final workings of trust, in human or machine agents. They do, however, provide an agent with a useful tool for the evaluation of situations and potential situations and cooperative relationships. In order to illustrate this further, this section provides two examples of trust in action using the above formulae. The situations are adapted from Connah and Wavish, 1990 [7], and represent a problem involving furniture removal. In the first situation, only two agents are present, and there are two pieces of furniture to be moved. In the second, there are three agents and three pieces of furniture. Each agent in both situations has the goal of moving a piece of furniture (different from any other agent) to the door. Unfortunately, they cannot lift the pieces by themselves, and thus need help. A diagram of the start state for each situation is given below. For more information and an introduction to the problem, see [7].

8.3.1 Example 1: Two Agents

In this example (see Figure 1) we have two agents, each of which has the goal of moving a piece of furniture to the door.

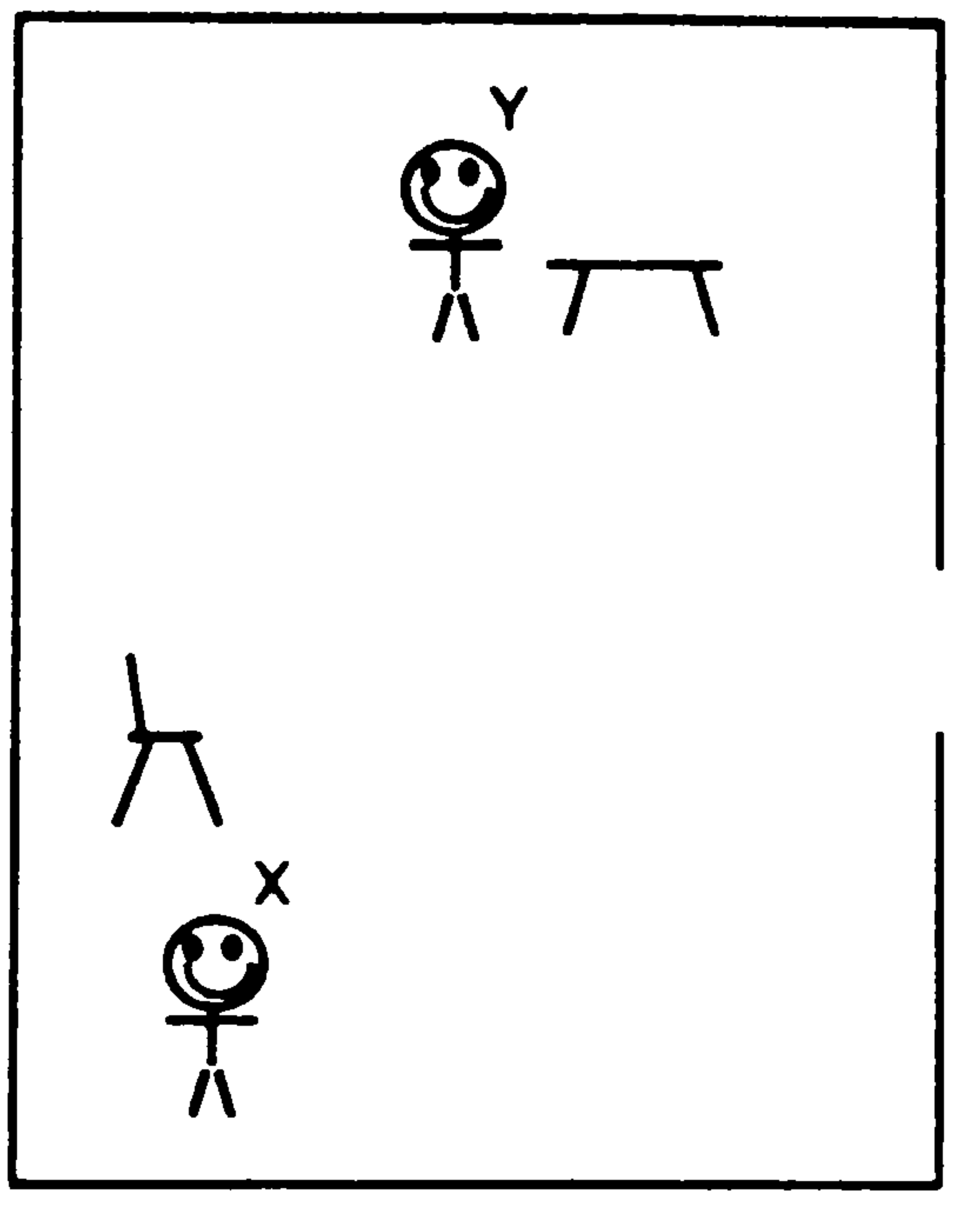

Figure 1: Furniture Moving — Two Agents at the Start of the Example

Each piece of furniture is too heavy to lift for an agent alone. In other words, to achieve their goals, the agents must cooperate. This situation is therefore an example of agents not having any choice about cooperation. They must cooperate. Some interesting points, however, will hopefully arise. Table 2 shows the workings of the first agent (call it x) with regard to how much trust it has in the second agent (y). In this situation, x decides it can trust y and will cooperate in the situation. Interesting results can be obtained with different starting values for trust, costs, benefits, and situational importance. Consider, for example, a situation where the cooperation threshold is higher than the amount of trust x has in y. Unfortunately, although this should lead to x not cooperating at all, cooperation is necessary if the goal is to be fulfilled. Cooperation will therefore have to be entered into. The result of a high threshold, however, would allow x to monitor y's behaviour very closely, and protect itself, for example by insisting y help move x's furniture first, thus guaranteeing the result it wants.

8.3.2 Example 2: Three Agents

The above example illustrates how an agent can use trust to reason about relationships with little or no knowledge about the other agents concerned (consider if x didn't know y at all — she would

Agent x trusts agent y by 0.85. She considers the situation to have an importance of 0.75. The Costs of the furniture not being moved amount, as far as x can estimate, to 0.60, this is quite high, but the benefits of moving the furniture are estimated at 0.70 — higher than the costs of them not being moved. $T_{x\,y} = 0.85$ $I_x^{\alpha_x} = 0.75$ $C_x^{\alpha_x} = 0.60$, $B_x^{\alpha_x} = 0.70$ The estimates of reciprocal trust are set, after careful consideration on the part of x, to be $\text{Est}_x(T_{y\,x}) = 0.65$ and $\text{Est}_x(\text{Est}_y(T_{x\,y})) = 0.70$.
From these values, x can work out the amount of situational trust in, or reliance on, y in this situation. From equation 1: $T_{x\,y}^{\alpha_x} = 0.78$
The risks associated with the situation (see equation 5) work out as follows: $\text{Perceived\_Risk}_x(\alpha_x) = \frac{0.6}{0.70} = 0.857$ This seems relatively high, but the competence of the other agent in this situation is not known, neither is anything known about the situation from previous experiences.
Since nothing is known with respect to the situation in hand and the agents present, competence is thus taken as the trust x has in y (from equation 4): $\text{Perceived\_Competence}_x(y, \alpha_x) = T_{x\,y} = 0.85$ From Risk and Competence, the threshold of cooperation can be found (equation 2): $\text{Cooperation\_Threshold}_x(y, \alpha_x) = 0.756$ The trust x has in y in that situation works out as 0.78. As this is greater than the threshold, x decides that it is safe enough to cooperate with y in moving furniture.

Table 2: Agent x's 'thoughts' on trusting agent y

just use the value of her general trusting disposition, T_x, in the place of $T_{x\,y}$). It also shows how some situations leave agents with little choice about cooperation — it is virtually coerced. In this case, perhaps the politics of coercion, rather than of trust, would be more applicable. The concept of trust becomes more interesting in a collection of agents, where different agents can be considered in the light of experience, competence, and so forth.

The present example considers the same situation, of furniture moving, but with three agents, and three pieces of furniture (see Figure 2).

We will consider the situation from x's point of view again. An interesting twist to the situation will be introduced in the fact that agent z is known and little trusted in general by x, but x 'knows' that z is, by chance, a specialised furniture removal agent, and thus extremely competent in this situation. Table 3 shows the 'thoughts' of agent x on cooperating with z.

The results of the deliberations are that x will favour working with y in this situation, despite z being a professional furniture mover. It is hoped that this mirrors real life situations to some extent — we may trust our best friend to help us move furniture rather than a professional, who may, perhaps, have less respect for our items of furniture.

9 Implementation

What has been presented thus far is a view of how trust can be represented in a fairly simple logical manner, but allowing for further expansion. It would be useful to discuss how a concrete implementation of such an idea could work. Some simple implementations have been arrived at, and it is envisaged that a more complex visual example will be forthcoming in the near future. The notation presented above is intended to make it easier to envisage the notion of trust working

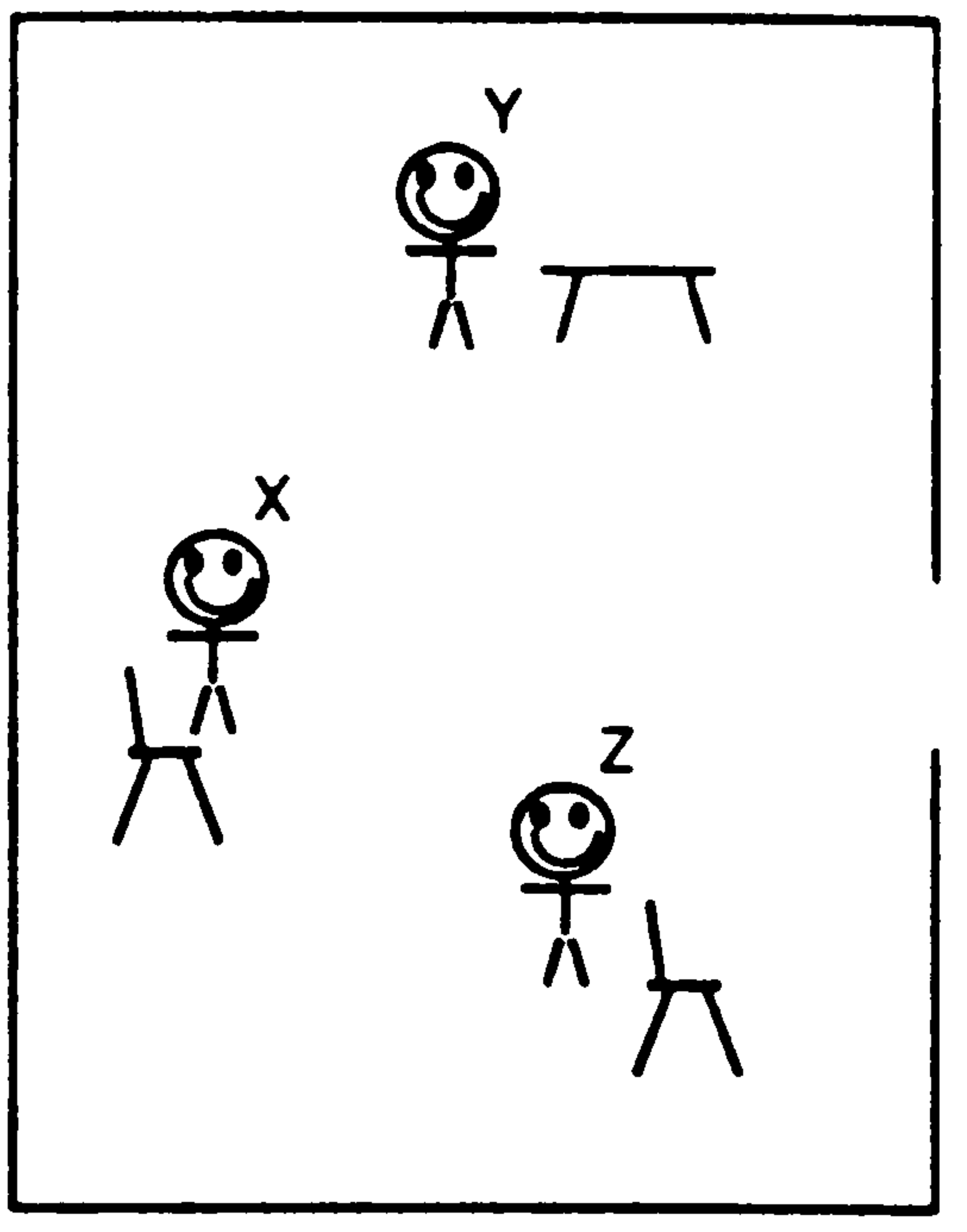

Figure 2: Furniture Moving — Three Agents at the Start of the Example

within an artificial agent, and in this section, some of the considerations involved in implementing the notion are briefly discussed.

Each new situation requires a re-evaluation, partial at least, of the amount of trust we have in the agent(s) with whom we are interacting. Clearly, the amount of trust we already have in the other agent(s) will play a large role in the judgements we make, indeed, should a similar situation have been carried through in the past with the same agent(s), our judgements are virtually made for us. Theoretically at least, it would be useful for our agents to have a knowledge of previous encounters with other agents *and* the situations, or context, of those interactions. In practice this may not be possible for reasons of space and speed, and so some other way of judging situations must be found. What should be remembered is a value relating to how we trust each different agent we have interacted with, which is alterable through our continuing experiences with these agents. The very least we can manage with is a general "trust" value which is altered with experiences with all agents, and applies to all agents.

In order to use a concept of another agent behaving 'well' or 'badly' in interactions, it is necessary to have some form of self-reference, such that the agent will know what it would have done in such a situation, and use this as a reference point for judging other agents. This self-reference would need to be embodied in our agents. An implementation of this could be based on finite state automata representations within the trusting agent. The trusting agent can attempt to model the behaviour of the trusted agent(s) using a simple finite state machine. The concept of behaving well or not then emerges simply from whether of not actual behaviour corresponds to the finite state machine represented by the trusting agent. This is an area for further consideration.

From the brief discussion above, it can be seen that the idea of trust is not task-specific, in other words it could be seen as an addition to what has been implemented thus far in any agent, and this late addition would not detract from either the workings of trust or the original task of the agent concerned.

A major task involved in the implementation is that of determining the workings of the various functions for altering the trust values discussed above. What has only been mentioned in passing thus far is one of the major aspects of trust: it is a *dynamic* reference to, in this example, interactions. The point is, it increases or decreases according to the behaviour of the trusted agents. For example, should y not help x after x had helped y move his piece of furniture, then the trust x has in y would decrease. Likewise, if y had helped, x's trust in y may increase. In addition, trustworthy or untrustworthy behaviour would alter the basic trust value of x (i.e. T_x) although not by as much (as was mentioned above, T_x is less dynamic than, for example, $T_x(y)$).

Whilst the concept of increasing or decreasing trust is clear, it is necessary to take care regarding

Agent x trusts agent z by 0.50. The costs and benefits remain the same as in the above example. (Table 2) $T_{x.z} = 0.50$ $I_x^{\alpha_x} = 0.75$ $C_x^{\alpha_x} = 0.60$, $B_x^{\alpha_x} = 0.70$ Now x perceives the reciprocal trust as lower than before: $\text{Est}_x(T_{y.x}) = 0.45$, and $\text{Est}_x(\text{Est}_y(T_{x.y})) = 0$ (x figures z doesn't care whether or not x trusts z). From equation 1, x works out the amount of situational trust in z in this situation: $T_{x.z}^{\alpha_x} = 0.24$ This, compared to the situational trust in y of 0.78 (above), is very low. However, we must note that z is a furniture moving agent, and as such, although the risks involved in the situation stay the same, the competence will change dramatically. The risks associated with the situation are as follows: $\text{Perceived\_Risk}_x(\alpha_x) = \frac{0.6}{0.70} = 0.857$ Competence is taken as a fairly high figure, for this example. x decides z is 100% competent at moving furniture (z has the certificate to prove it!): $\text{Perceived\_Competence}_x(y, \alpha_x) = 1.00$ From Risk and Competence, the threshold of cooperation can be found (equation 2): $\text{Cooperation\_Threshold}_x(y, \alpha_x) = 0.64$ Whilst this is lower than the Cooperation_Threshold for agent y in the same situation, it is still much higher than the situational trust x has in z, of just 0.24. In this case, then, x would decide not to cooperate with z if there was a choice (there is — y). In a situation with no choice, a substantial safety net could be initiated (e.g. take out insurance, don't pay until the job is done, etc. The safety net is largely dependent on the situation at hand).

Table 3: Agent x's 'thoughts' on trusting agent z.

the sensitivity of the trust values. Too sensitive, and the agents in question would never achieve a stable relationship; too static, and the agents concerned may well be not trusting enough or may be too easy to take advantage of. The ideal sensitivity will be arrived at through experimentation, and may well be task specific, contrary to the general idea of trust itself. In addition, the measure of sensitivity allows some control over the trusting behaviour of the agent.

10 A Partial Order

Since we have introduced values to the notion of trust, it may be necessary to say a few words about how those values compare to each other. In other words, how is it possible to order the different trust values, and how does trust 'distribute' (wrong word?) through agents.

Trust using values is a partial order. In other words, it is not possible to say that x trusts z more than y, even if $T_{x.z} = 0.8$ and $T_{y.z} = 0.7$. There is no order there. This is because trust is such a subjective notion. What x calls 0.8, y may call 0.5.

What implications does this have? It suggests firstly that trust is not a transitive relationship. From the point of view of an agent who knows some agents and not others, this means that estimating original trust in agents unknown to himself is difficult. It was originally envisaged that such an estimate could be based precisely on the views of other agents, if there were any[6]. This is still the case, but more thought is necessary on the part of the trusting agent. It is suggested that, for the present, the following rule should apply (assume the situation where x knows and trusts y, but not z, and y knows and has a trust value for z):

$$T_{x.z} = T_{y.z} \times T_{x.y} \tag{6}$$

[6] One of the benefits of a formalism using values is that such things are possible.

Note that, if one or both of these values are negative, the result should be set negative, no matter what sign it has.

11 Harmony — Wa and Trust

The environment in which our agents will eventually exist within may not be a friendly one. Indeed, that is one of the reasons for introducing trust in the first place, as was discussed above. Trust, however, could play a large role in a society in which harmony existed. The concept of harmony, or *Wa*, is discussed in [36], and is being taken as an 'ideal' for interactions between agents in a multi-agent system in [21]. Briefly, the intention is to have a society where harmony exists, and cooperation and social behaviour is thus expected. In this situation, trust is necessary because of the need to trust everyone to behave in that way. A further extension to this idea would encoporate some of Minsky's ideas with respect to law-governed systems [24, 23], thus removing much of the 'danger' inherent in trusting unkown agents — the safety net is provided in the form of the laws of society. This is work still to be investigated.

12 The Limitations of Trust

The concept of trust introduced in this paper has its limitations. Not least is that there is a considerable amount of work still to be done. For example, thus far we have considered only one of the many different areas where trust can prove useful, namely that of initiating, or considering, cooperation with other agents. Whilst this is useful, it is only one area. Other important areas were briefly discussed above, notably the use of trust for gauging confidence in the information given to an agent, as in source control [10].

In addition, trust can help in the formation of groups: "Perhaps there is no single variable which so thoroughly influences interpersonal and group behaviour as does trust..." [12, page 131]. The formation of a group is basically an extended cooperative relationship between more than two agents. As such, the same kind of trust may well be used in forming the group. A question may arise in how groups themselves trust. For example, does a group trust another group more or less than another individual. Does a group see itself as a single entity, or just many individuals working to a common purpose? All these questions have yet to be addressed.

Trust provides an agent with a tool for judging the future, based on experience of the past. It is, however, limited when applied alone. It is not envisaged that an agent use trust as the sole measure of certainty in situations, much less as the only means of making a decision. The examples given above are intended to illustrate how trust can be used in such situations, but they are limited, as might be expected. When coupled with other methods, such as utility theory or theories of rational behaviour (e.g. [29]), trust provides an agent with a powerful and useful tool when interacting with other agents.

The heuristics given above present their own considerations. One of the problems with attaching values to things of that nature appears to be the different interpretations that can come from the values. It is hoped that the implementation of a trusting agent will enable a thorough investigation of these heuristics. They are, as such, introductory, but solid in that they work in a way in which trust would be expected to act in an agent, and allow for experimentation and discussion.

There are also problems with deadlock. Consider an agent in the furniture moving example who need help, but the only other agent around refuses to give it, and needs no help for itself (no furniture to move?). Nothing will happen — a classic deadlock situation, it seems. This suggests that such situations need reciprocity. This is, however, not so much a problem with trust as with cooperative situations in general. Nevertheless, it is a problem which will need addressing.

Finally, there is a large amount of complexity involved for one agent when considering many other in terms of trust. This may or may not be unavoidable, but it does suggest that trust may not be of much help in situated agents, where the behaviour is dictated by the environment. It will, however, help an agent when considering the environment and possible environment states, as briefly mentioned above.

13 Conclusions and Ideas for Further Work

The notion of trust presented here is extendable. Further work will concern the decisions an agent makes with regard to its environment; how the agent trusts what it interacts with in a physical sense. Secondly, the fact that an agent trusts another does not impel the trusted agent to behave in a 'proper', trustworthy manner, although the fact that the other agent is trusting implies an acceptance of this 'danger'. As discussed briefly above, what would be useful is to include in the trusting world a law of some sort, not unlike the law-governed systems of Minsky [24, 23]. These systems provide laws which "...affect only what an agent may be able to do, not what it actually will do." [24, page 291]. What this means to the trusting agent is that the agent being trusted may be malevolent, but not capable of being malicious, and of actually damaging the trusting agent. In such a situation, trust may be built up more easily and safely, and this will be further studied in the future. In addition to the work discussed in section 12, further work will investigate the possibilities of using the concept of trust in, for example, distributed operating systems, as an analysis tool for more complex multi-system interactions, and so forth. Finally, the use of trust as a tool for the reduction of complexity, as briefly mentioned above, is an interesting avenue of work in itself.

In conclusion, a concept of trust has been introduced for agent interactions. The benefits of the concept include the allowance of a dynamic reference to interactions between agents, both with reference to the agent itself and to interactions with other agents. It allows at least minimal interactions and cooperation between agents where ordinarily there would be none. Most importantly the agents concerned are making an implicit acknowledgement of the possibility of malevolence or mistakes on behalf of the other agents, and as such there can be some form of backup against such occurences. It is thus of benefit both in that interactions are allowed, but with an acknowledgement of the problems involved. In the near future a working implementation of such a trusting agent will be forthcoming.

Acknowledgements

Many thanks to Ian Wilson, Harold Thimbleby, Lynne Coventry, Tom Kane, Andy Cockburn and others for constructive comments on earlier drafts of this paper. Also to the attendees of the MAAMAW'92 workshop in Italy for instructive insights into some of the aspects of trust. The comments from anonymous referees enabled the author to improve this paper a great deal, for which, many thanks. The author is a SERC CASE student, partially supported by Canon Research Europe.

References

[1] Robert Axelrod. The evolution of strategies in the prisoner's dilemma. In Lawrence Davis, editor, *Genetic Algorithms and Simulated Annealing*, pages 32–41. Pitman, London, 1987.

[2] Robert Axelrod. *The Evolution of Cooperation*. Penguin Books, London, 1990.

[3] Roy L. Behr. Nice guys finish last – sometimes. *Journal of Conflict Resolution*, 25(2):289–300, June 1981.

[4] Richard Boyle and Phillip Bonacich. The development of trust and mistrust in mixed-motive games. *Sociometry*, 33:123–139, 1970.

[5] Stephanie Cammarata, David McArthur, and Randall Steeb. Strategies of cooperation in distributed problem solving. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1983.

[6] Man Kit Chang and Carson C. Woo. SANP: A communication level protocol for negotiations. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in an Artificial World, Germany*, 1991.

[7] David Connah and Peter Wavish. An experiment in cooperation. In Yves Demazeau and Jean-Pierre Muller, editors, *Decentralized AI*, pages 197–212. Elsevier Science Publishers (North-Holland), 1990.

[8] Peter A. Danielson. Artificial Morality: Prolog and the Prisoner's Dilemma. Corrected version of a paper presented at the Fifth International Conference on Computers and Philosophy, Stanford University, 8–11th August, 1990., September, 1990.

[9] Morton Deutsch. Cooperation and trust: Some theoretical notes. In M. R. Jones, editor, *Nebraska Symposium on Motivation*. Nebraska University Press, 1962.

[10] Roberto Garigliano, Albert Bokma, and Derek Long. A model for learning by source control. In Bouchon, Saiger, and Yager, editors, *Uncertainty and Intelligent Systems*, pages 163–170. Springer Verlag, Lecture Notes in Computer Science, LNC 313, 1988.

[11] Matthew L. Ginsberg. Decision procedures. In M. Huhns, editor, *Distributed Artificial Intelligence, Volume 1*, chapter 1, pages 3–28. Pitman, London, 1987.

[12] Robert T. Golembiewski and Mark McConkie. The centrality of interpersonal trust in group processes. In Cary L. Cooper, editor, *Theories of Group Processes*, pages Chapter 7, pages 131–185. Wiley, 1975.

[13] Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, July 1990.

[14] Lars Herzberg. On the attitude of trust. *Inquiry*, 31(3):307–322, September 1988.

[15] Nick R. Jennings. On being responsible. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in an Artificial World, Germany, Panel Session*, 1991.

[16] John T. Kohl. The evolution of the Kerberos authentication service. In *Europen '91 - European Conference on Open Systems*, pages 295–313, 1991.

[17] Olli Lagenspetz. Legitimacy and trust. *Philosophical Investigations*, 15(1):1–21, January 1992.

[18] Bernhardt Lieberman. i-Trust: a notion of trust in three-person games and international affairs. *Journal of Conflict Resolution*, 8(3):271–280, 1964.

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

[20] Stephen Marsh. 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*, 1992.

[21] Stephen Marsh and Harold Thimbleby. Wa — Harmony in DAI Environments. In preparation, 1992.

[22] Steven McNeel. Training cooperation in the prisoner's dilemma. *Journal of Experimental Social Psychology*, 9:335–348, 1973.

[23] Naftaly H. Minsky. The imposition of protocols over open distributed systems. *IEEE Trans. Software Engineering*, pages 183–195, February 1991.

[24] Naftaly H. Minsky. Law-governed systems. *Software Engineering Journal*, pages 285–302, September 1991.

[25] Bonnie M. Muir. Trust between humans and machines, and the design of decision systems. *International Journal of Man-Machine Studies*, 27(5 & 6):527–539, Nov/Dec 1987.

[26] Chisato Numaoka. Conversation for organisational models. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in an Artificial World, Germany, Panel Session*, 1991.

[27] Derek Parfit. Prudence, morality, and the prisoner's dilemma. In Jon Elster, editor, *Rational Choice*, pages 34–59, 1986.

[28] Julia Rose Galliers. A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict. Technical Report No. 172, University of Cambridge Computer Laboratory, 1989.

[29] Jeffrey S. Rosenschein. *Rational Interaction: Cooperation among Intelligent Agents*. PhD thesis, Stanford University, October, 1985.

[30] John F. Shoch and Jon A. Hupp. The "worm" programs: early experience with a distributed computation. *Communications of the ACM*, 25(3):172–180, March 1982.

[31] Eugene H. Spafford. The Internet Worm: Crisis and Aftermath. *Comms. ACM*, 32(6):678–687, June 1989.

[32] Robert L. Swinth. The establishment of the trust relationship. *Journal of Conflict Resolution*, 11(3):335–344, 1967.

[33] Harold Thimbleby. Can viruses ever be useful? *Computers and Security*, 10:111–114, 1991.

[34] Donnell Wallace and Paul Rothaus. Communication, Group Loyalty, and Trust in the PD Game. *Journal of Conflict Resolution*, 13(3):370–380, 1969.

[35] Thomas Y. C. Woo and Simon S. Lam. Authentication for distributed systems. *IEEE Computer*, pages 39–52, January 1992.

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

Trust in Distributed Artificial Intelligence

Stephen Marsh

Department of Computing Science and Mathematics, University of Stirling
Stirling, FK9 4LA, SCOTLAND
email: spm@cs.stir.ac.uk, Telephone: (+44) 786 467444

Abstract. A discussion of trust is presented which focuses on multi-agent systems, from the point of view of one agent in a system. The roles trust plays in various forms of interaction are considered, with the view that trust allows interactions between agents where there may have been no effective interaction possible before trust. Trust allows parties to acknowledge that, whilst there is a risk in relationships with potentially malevolent agents, some form of interaction may produce benefits, where no interaction at all may not. In addition, accepting the risk allows the trusting agent to prepare itself for possibly irresponsible or untrustworthy behaviour, thus minimizing the potential damage caused. A formalism is introduced to clarify these notions, and to permit computer simulations. An important contribution of this work is that the formalism is not all-encompassing: there are some notions of trust that are excluded. What it describes is a specific view of trust.

1 Introduction

This work introduces the concept of trust and its uses in Distributed Artificial Intelligence (DAI). We first present a discussion of trust and some of its forms. We then follow with a formalism which embodies the concept of trust, and facilitates decisions regarding trust. The advantages of this are clear — the formalism allows clarification of the concept; it allows experiments to be carried out with trust; and, due to its implicit simplicity, it allows a fuller understanding of trust to be reached in a stepwise fashion.

Trust is a value-laden concept, which everyone knows of and uses, most of the time [25]. The notation which will be presented here is not intended to be a definitive view of the way trust and reliance work (which is, in any case, probably not possible). It provides a focussed discussion article and is intended to initiate interesting, productive debates on the topic. If the discussion and the notation are too weak for some, they have an invitation to expand or contract it, and to let me know of their thoughts and comments. The formalism proposed here is in no way final: it is an experimental piece, and by its nature continuously in flux. It applies to a cooperative situation between two or more agents from the point of view of one of the agents. For situations involving other aspects of cooperation or areas where trust is an issue, the formulae may well need adjustment. The intention is to prove that such considerations are indeed possible with trust, and that it provides a helpful addition to the decisive powers of agents. Further discussion can be found in the author's thesis [28].

2 What Trust is

Definitions of trust are rare. Much of the literature pertaining to trust mentions the word, or the concept, with no real definition, and the resultant confusion over the concept does little to help us understand it [4]. Despite the fact that trust is "a basic fact of human life" [3], we still stumble over definitions.[1] A basic fact which is accepted is that trust implies a risk of some sort, "One trusts when one has much to lose and little to gain" [12, page 304]. From many of the definitions, this is taken to be the case, as in Swinth [37], who takes this point of view, stating that entering a trusting relationship is "choosing to take an ambiguous path that can lead to a beneficial event or a harmful event depending on the behaviour of the other person — where the harmful event is more punishing than the beneficial event is rewarding."

In our work, we take a less extreme view. The ambiguous path still exists, but we do not assume that the harmful effects outweigh the positive ones. Deutsch states that, where positive effects are greater than negative, the choice of the ambiguous path is not trust, but a gamble (see Deutsch, 1973 [13]). It is possible that this is the case. This, however, does not affect the validity of the formulæ presented here, and Deutsch himself presents gambling as a form of trust (see Deutsch, 1973). Our definition of trust, then, follows Deutsch's up to this point. For clarification, Deutsch's definition is:

> (a) the individual is confronted with an ambiguous path, a path that can lead to an event perceived to be beneficial (Va+) or to an event perceived to be harmful (Va−); (b) he perceives that the occurence of Va+ or Va− is contingent on the behaviour of another person; and (c) he perceives the strength of Va− to be greater than the strength of Va+. If he chooses to take an ambiguous path with such properties, I shall say he makes a trusting choice; if he chooses not to take the path, he makes a distrustful choice.
>
> Deutsch, 1962, Page 303.

3 Trust *vis a vis* DAI

Implicit in the notion of DAI is the concept of decentralisation. Since decentralisation implies a lack of central control, and with it a lack of explicit guidance in the 'right' direction, it becomes necessary, in order to carry through successful interactions with other agents, to develop some judgement as to the worth of these interactions and the risk associated with them. The concept of trust has already been widely field-tested with respect to the human race [6, 12, 37, 40]. It provides us with an ideal measure of expectation of risk, and since the risk is implicitly acknowledged in the form of a trusting relationship then some form of preparation can be made against untrustworthy behaviour, which might otherwise be fairly damaging — a form of 'safety net'.

[1] This, again, is where the formalism can help, in clarifying the situation.

In a laboratory, we, and our agents, are acting under controlled conditions with the knowledge of what kind of behaviour to expect, and this knowledge comes forward and is instilled, unconsciously or not, in our experiments. In the case of DAI, the agents that are built are as subject to this 'rule' as anything else: "it makes little sense to ask *why* they are helping one another; they help each other because they have been designed that way" [34, page 12]. What the concept of trust should help ensure is that our agents are more robust with respect to interactions with agents that are not our own, and interactions of a type that is not forseen. In addition, the concept of trust in certain future eventualities and in the expected behaviour of the environment (including other agents) functions as a tool for the reduction of complexity for the trusting agent [25, 26].

Assuming blindly that cooperation is a good thing is not necessarily the correct approach, although this depends on the viewpoint of our agents. Behr [5] points out that even something as conceptually simple as the Prisoner's Dilemma [1, 2, 29, 33] can be viewed from differing perspectives, depending on whether one wishes to score highly, as Axelrod had assumed [2], or whether one wishes to defeat the opponent. In the latter case, the more successful strategies were not so nice, since to win, one must "defect more than one's opponent does" [5]. The moral of this exercise, if there is one, is that if we assume that everyone is out to cooperate, we are mistaken. Lack of cooperation may actually benefit others. What is perhaps better to assume is a notion of self-interested agents, all out to get the best they can [24] in so far as they may see it. In some cases that means that they cooperate, in others, their behaviour can range from non-cooperation to downright maliciousness. Since we cannot presume to know what their behaviour may be at any given future time, the notion of trust put forward in this paper relies on a judgement based on experience, coupled with, if available, past knowledge of the agent to be trusted and their behaviour.

3.1 Why Trust?

DAI is concerned primarily with cooperation: why it happens, who cooperates with whom, when, and to what extent. However, DAI has thus far concentrated on why cooperation takes place, why it is useful to an agent to cooperate, with little thought being given to any of the other aspects of cooperation. In particular, little has been done on agents surviving in the 'outside world', as opposed to in restricted experimental worlds.

Cooperation is generally accepted as being a good thing [2, 8, 10]; however, there are situations where agents have conflicts of interest, and some form of coping with these conflicts is required [14, 34]. Whilst some studies take coping into account, [9, 14, 34], they make the important assumption that the agents are *trustworthy*. This is indeed seen as "absolutely essential" [34], at least in a situation where the agents are communicating, and may well be more important if communication is not allowed. The agents that we design for use within our experimental worlds will most likely not be robust enough to survive outside the laboratory. Why is this? Altruism, however desirable, is not the "name of the game" in the world we live in. Indeed, taking the world of computing alone, malevolence, not altruism, appears to be prevalent [38], and this is irrespective of how well-intentioned the work may be to start with (Shoch and Hupp [35], for example, give details of the idea of a useful "worm" program, the basic premise of which was used to devastating effect not so long ago [36], despite never being intentionally designed for malevolent purposes).

Taking this lack of altruism into account, some measures must be taken to make our agents less vulnerable to others' incompetent or malevolent behaviour. There are different approaches to this. The first, most obvious, is not to interact with anyone we don't know, and it is this method which is used in the simple username, password schemes on computers, and other security systems such as Kerberos [22]. This approach is too restrictive to allow a normal interaction between agents. Another method is to ignore the vulnerability of our agents, and hope they will not be exploited. This, too, is shortsighted. This work explores the idea that there is a continuum between these two extremes which can be used to the advantage of agents, and that the points on this continuum represent varying degrees of trust on behalf of an agent.

4 Forms and Issues of Trust

4.1 Reliance

Different situations require different forms of action. While I may trust you to drive me to an airport, trusting you to fly the plane is another matter! What this implies is that for different interactions between agents, a different kind, or form, of trust may be required, in that different things need to be taken into account with regard to different situations. Note that this is not the same as having different *degrees* of trust in an agent. Indeed, there can be different degrees of the different *kinds* of trust. This is closely coupled to the agent's view of a situation (see below).

The amount of trust in a person or agent does not change from situation to situation solely because the situation changes; rather, the reliance, based on the trust in the agent, changes according to the different situation. Hence, if I were to speak in terms of percentages of trust or reliance, I may say I trust (rely on) you 50% to drive me to the airport, but only 20% to fly the plane. The amount of reliance I may have in you at a given moment, or in a given situation, is a function of the amount of trust I have in you in general, in addition to my experience of your actions in similar situations in the past, and the competence I perceive you to have in the situation. In some situations, however, the *value* we place on trust is higher than in others. As an example, if I were in a situation where a wrong decision could cost me my life, I would think very hard about taking your advice, especially if I didn't trust you completely.

4.2 Security

Between blind trust and complete mistrust lies a continuum of varying degrees of trust. The security measures of today offer a mixture of both blind trust and complete mistrust. As an example, if I am a complete stranger attempting to access a computer, if I don't know a password, I will not be allowed access. If, however, I stumble across a password, or obtain another method of entry, then the computer affords me almost complete trust, limited only by the prior actions of the system administrator. The same applies to agents — they may be initially completely non-trusting, but once the outer defences are penetrated, they can be used in any way, to transmit viruses, for example.

The concept of dynamic trust enables an agent or system to interact (at least to some extent) with some other agent(s). The limits of interaction change with experience of the action of the other agent(s). In other words, an agent can interact with others to a certain extent, trusting them, or relying on them, to that extent. In the light of trustworthy behaviour on their part, the extent to which the agent trusts them (and hence to which it is prepared to interact with them) will increase, and untrustworthy behaviour results in a decrease in the amount of trust.

4.3 Cooperation

The advantages of cooperation are many and diverse. Indeed, many tasks cannot be performed by one agent alone, perhaps because the agent does not have all of the knowledge necessary to formulate a solution to the problem, perhaps because the agent cannot physically perform a task without help [10], and so forth. In its simplest form, a cooperative relationship is between two agents. In order to initiate a cooperative relationship, some form of trust is required, "the initiation of cooperation requires trust whenever the individual, by his choice to cooperate, places his fate partly in the hands of others."[12, page 302]. The amount of trust, and the importance of it, depends on the situation the relationship is formed to handle. Thus far in DAI, what has been assumed is that is all situations, trust is 100% (which we disallow in any case - see below).

4.4 Information Sharing and Gathering

Why would we want to trust a stranger? The answer lies in what we are expecting. From any relationship, there is some way of benefiting for both, or all, of the members of the relationship. In information-sharing, for example, all information may well be useful, and finding out which is useful and which is not depends on having the information to hand. If I were not to trust ninety percent of people, I may well lose what I would consider vital information.

Consider an agent seeking information in the world at large. As such, this agent may well be given information by other, unknown agents. This information has to be judged in some way, and the notion of trust may play a part here. In other words, an agent can use a measure of trust in the agent giving the information, and also the validity of the information concerned. This latter measure could be based on the prior knowledge of the agent, and the way in which the information correlates with what is already known [7]. In this case, a knowledge of or trust in the other agent may not be necessary; for example, if I were to show you a white piece of paper and tell you it was white, all you would need to trust was your own judgement. Some information may also be this self-evident. Using already known and accepted knowledge may not be possible, since nothing may be known about the information in question, and thus there is a useful 'backup' with the measure of trust in the agent giving the information. For example, were I to put a piece of paper behind a chair where you could not see it, and tell you it was white, you would have to trust me in order to believe this information in that situation. Trust would allow you to accept this with some measure of certainty or uncertainty. Indeed, the work of Garigliano *et al* in the area of Source Control is an example of this to some extent [16].

4.5 Trust as Reduction of Complexity

The concept of trust acts as a tool for the reduction of complexity for an agent as regards future eventualities [25]. That is to say, an agent need only consider trusted eventualities, or world states that arise from trusted actions and world states. In other words, we do not have to consider all possibilities if we trust that some will not happen. That does not, however, mean that we will not prepare for them. We can still take precautions, even against something we know nothing of. The result of this is that the amount of processing devoted to considering future events is likely to be drastically reduced. This would be a great help for an agent as it makes a plan. At present, the idea is very much in its early stages. It is hoped, however, that an agent architecture exploiting trust as a tool for the reduction of social and environmental complexity can be implemented.

5 A Formalism

We have proposed areas of work where a consideration of trust is of use. The paper now introduces a formalism for trust which allows a DAI agent to reason with and about trust. The justifications for such a formalism are, at the least, threefold.

- A formalism can help to avoid ambiguities in discussing a concept. For a concept as value-laden as is trust, the formalism can help avoid many possible disagreements by making clear what is actually being discussed.
- If we are to progress to actually implementing some kind of trust apparatus within an agent, the formalism can help in that implementation, since it may be a small step from there to a working example.
- The possibility of having worked examples can help us to determine whether or not the proposed heuristics actually work. For example whether incrementing trust by such and such an amount following trustworthy behaviour is justified, or if the method of using trust in decisions is satisfactory.

With a formalism and some heuristic formulae, it is possible to determine experimentally and theoretically the answers to these and other questions.

6 Notational Definitions

6.1 The Agents

We represent particular agents by the letters a to z. Each agent is a member of $\mathcal{A}$, the set of all agents. An agent can be considered to be an independent entity extant in a world populated with other such entities, each of more or less complexity than itself. In the introductory notation presented below, we present each formula unrelated to any other.

6.2 Situations Agents Find Themselves In

A situation is defined as a point in time relative to a specific agent. This allows us to consider two or more agents, at the same point in time, and even in the same place, to be in different situations, since each agent may have different knowledge, beliefs, or intentions at that time, colouring their perception of the situation. For example, consider the example given by Halpern [19], where there are several children playing, some of whom have mud on their foreheads. For each child, the situation is different, since they do not know if they have mud on their forehead (they may have), but they can see all the other children. In a situation where knowledge may be partial, that situation is seen differently by all in it. So in our notation situations are represented as the greek letters, α to ω, with a suffix representing the agent concerned (the agent who sees that situation from her point of view), so we have α_x to ω_x for situations from the point of view of agent x.

6.3 A Basic Trust

An agent, as a trusting entity, has a basic trust 'value', derived from previous experience. This value, which is dynamically altered in the light of all experience, is used a great deal in the formation of trusting relationships with unknown agents, and is represented as T_x, and has a range over $[-1, 1)$, taking values of $-1 \leq T_x < +1$. A value of -1 represents total distrust, 0 is not trusting at all.[2] A value of +1 indicates total trust. This is what is sometimes called 'blind' trust. It is disallowed here, since 'blind' trust implies a lack of consideration, or careful thought, on the part of the truster. This is commonly perceived to be a bad thing [11], and indeed is not accepted as 'real' trust by many philosophers [7]. It is important to realise that T_x does not correspond to the amount of trust x has in any specific agent, but only to the general trusting 'disposition' of x. Since it represents this basic trust, it is less 'fluid' and changeable than a trust in any specific agent.

[2] Note that these are not the same. Not trusting someone is not the same as having no opinion on the matter. Put another way, (distrust x) $\not\equiv$ (have no trust in x).

6.4 The Trust Value

Given two agents, $x, y \in \mathcal{A}$, to say x trusts y, we write: $T_x(y)$. In addition to being a representation of the fact that x trusts y, this is a value, over $[-1, 1)$, of the amount of trust x has in y.[3] In other words, should $T_x(y) = 0.5$, then x has a trust in y of 50%. By this, we mean that the chances of x taking Deutsch's ambiguous path are 50%. This is clarified later. The trust value is a view of a particular agent of another with regard to the trusted agents general capabilities. As was discussed briefly above, different situations may require different views of trust. The amount of trust in a particular agent in a given situation is represented here as, for example, $T_x(y, \alpha_x)$, over $[-1, 1)$ for x's *situational* trust in, or reliance on, y to perform correctly in situation α_x. Note the interchangeability of the concepts of situational trust and reliance here. At present, they are considered to be identical. However, it may be feasible in the future to separate the two concepts, although closely related, to reflect more closely the more philosophical views on the subject [20, 23].

6.5 Further Notation

The agent's estimate of how important a situation is to itself is a value over $[-1, 1]$, represented by $I_x(\alpha_x)$ here. The importance value in this example is x's estimate, or subjective measure, of how much situation α_x means to it. A value of 0 means that the situation is of no importance to agent x. The importance of a situation to an agent is useful in determining the amount of situational trust to place in an agent at any given time, as will be seen below; as an example, importance could be measured in terms of payoff functions [17, 34]. In the examples below, it will be introduced as an arbitrary value. Further models of trust will expand the idea of importance, and how it can be estimated by the agent.

Related to the importance of a situation are the concepts of costs and benefits pertaining to that situation. The costs of a situation are measured in terms of the problems associated with incompetent or malevolent behaviour on the part of another agent in a relationship. Initially, the agent can only estimate the potential costs of a situation, based on past experience of similar situations, or just a 'rule of thumb' if that situation is new to the agent. They are represented here by $C_x(\alpha_x)$, with a value over $[-1, 1]$. Note that any agent(s) involved in the situation are not represented. This is because the potential costs of a situation are relative only to the agent concerned. As such, whoever is to be trusted

[3] It is realised that attaching values to things creates its own problems, such as "what is the difference between a trust of 0.51 and a trust of 0.52?" This is acknowledged, but the values are at present retained for two reasons: Firstly, attaching a value to something allows us to talk in quantitative and qualitative terms, such as, "a trust of 0.75 is quite high, and higher than a trust of 0.70". Secondly, the use of values allows us to use calculations in arithmetic such as those introduced below. The difficulties remain, but it is hoped that the benefits outweigh these.

and cooperated with, the potential costs of untrustworthy behaviour remain the same.

The benefits of a situation, or at least the expected benefits of trustworthy behaviour from the agent(s) being worked with, play a large part in decisions of whether or not to cooperate in the first place. In the notation here, the benefits are represented as $B_x(\alpha_x)$, over $[-1, 1]$.

As was mentioned above, the costs and benefits of a situation are related to the importance of that situation to any agent. Importance goes further than a simple weighing up of costs and benefits, however, and may include some knowledge or assumption about future benefits, preparation for further cooperation, and so forth. The concept of utility, notated $U_x(\alpha_x)$ is also a little deeper than this. Utility implies a rational agent weighing up costs and benefits for a situation, and coming up with a single value which (in this case) is in the range $[-1, 1]$. For this example, we take utility to be as simple as this.

A summary of the notation used in this paper is presented in Table 1.

Situations are represented by $\alpha_x \ldots \omega_x$.
Individual agents are represented by $a \ldots z$, and are members of $\mathcal{A}$, the set of all agents.
Basic Trust Value for Agent x: T_x
General Trust x has in y: $T_x(y)$
Situational Trust (Reliance) x has in y in situation α_x: $T_x(y, \alpha_x)$
Importance of situation α_x to agent x: $I_x(\alpha_x)$
Potential costs to agent x following untrustworthy behaviour from another trusted agent in situation α_x: $C_x(\alpha_x)$
Potential benefits to agent x following trustworthy behaviour from another trusted agent in situation α_x: $B_x(\alpha_x)$
Utility of situation α_x for agent x: $U_x(\alpha_x)$

Table 1. Summary of notation for Trust — See the text for more details.

7 Rules for Interactions

Using the notation introduced above we can reason about trusting relationships. An example is given below of how trust can be used to determine the value of a cooperative relationship to a trusting agent. The examples presented below are intended to demonstrate how a particular agent can reason about the future using trust. The equations and the values associated with them are not intended to represent the cognitive working of trust in humans, and neither should they be taken as a final statement of how trust works within agents. They are intended to illustrate the thesis that a theory of trust, and approximate workings of that theory, can be embedded within an agent and used by that agent to help make decisions in particular situations.

7.1 Situational Trust

The trust an agent has in another in a specific situation is a function of the amount of trust in that agent in general and the importance of the situation to the trusting agent. In addition, the trust the other agent has in the first may play a part — knowing that you trust me may help me to reciprocate that trust, as does an estimate of how much I think you may trust me:[4]

$$T_x(y, \alpha_x) \; = \; f(\, T_x(y)\, ,\; I_x(\alpha_x)\, ,\; \widehat{T_y(x)}^{x}, \widehat{\widehat{T_x(y)}^{y}}^{x}\;)$$

Where $\widehat{T_y(x)}^{x}$ is an estimate made by x. Of course, x can only estimate the amount of trust y has in her ($\widehat{T_y(x)}^{x}$). This estimate may be wildly wrong, but could be quite sophisticated, taking into account what other agents may have said, previous behaviour of y, and so forth. Likewise, the estimate of expected trust , $\widehat{\widehat{T_x(y)}^{y}}^{x}$, is useful. The recursion here may cause problems of infinite regress. However, the extra information attained from each level of recursion becomes of less and less significance. It has thus been decided to limit the level of recursion. In the formulæ given below, these estimates are ignored, and the basic formulæ presented. They can then be added at some later time to ascertain their effect, if any, on agents' decisions.

We use the following formula to calculate the value of situational trust:

$$T_x(y, \alpha_x) = \widehat{T_x(y)}U_x(\alpha_x)I_x(\alpha_x) \tag{1}$$

Thus, informally, we define trust of an agent x of 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 [15].

The estimate of general trust, $\widehat{T_x(y)}$, is an interesting notion. There are various means that the agent can use to obtain this value; for example, it may be

[4] And so forth. This recursive nature of trust may be problematic. The final formula takes into account just one level of recursion. This is, of course, expandable, but a limit point may exist somewhere. See Dasgupta, 1990, page 51.

the maximum trust value the agent can remember, or the minimum (see Marsh, 1993 [27], Marsh, 1994 [28]). The most common estimate in our work thus far has been obtained by using the following, over a set, A, of tasks that x can 'remember':[5]

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

Inevitably, given the fact that utility, importance, and trust are fractional, the above equation results in a low estimate for situational trust. There are, of course, alternative equations. See the author's PhD thesis [28].

7.2 Cooperation Threshold

In order to cooperate with agent y, the trust x has in y for that particular situation has to be above a certain threshold value, itself a function of the importance, costs and benefits of the situation concerned:

$$\text{If } T_x(y, \alpha_x) \geq \text{Cooperation\_Threshold}_x(\alpha_x) \;\Rightarrow\; \text{Will\_Cooperate}_x(y, \alpha_x)$$

where we take: $\text{Cooperation\_Threshold}_x(\alpha_x)$ to be:

$$\frac{\text{Perceived\_Risk}_x(\alpha_x)}{(\text{Perceived\_Competence}_x(y, \alpha_x) + \widehat{T_x(y)})} \times I_x(\alpha_x) \tag{2}$$

Here, $\text{Perceived\_Competence}_x(y, \alpha_x)$ reflects what was discussed above, and allows for cooperation with an agent which is not trusted very much in general, or with an agent in an important situation, where that agent is known to be reliable and competent to a high standard in that or similar situations. The $\text{Perceived\_Risk}_x(\alpha_x)$ represents the agent's best estimate of the potential costs and benefits of the situation. Both of these are expanded as follows:

Perceived_Competence

1. Since competence takes into account the agent to be trusted, the Perceived Competence measure is based on experience in similar situations, experience of the same agent in similar situations, and knowledge of that agent's capabilities in similar situations. These considerations are naturally quite subjective. It is possible to visualise an agent keeping some form of database which records situation types, agents within those situations, and the observed competence of those agents in those situations. For the purposes of this example, however, we use the following estimate of competence.

[5] Bearing in mind that agents are likely to have a finite memory, A will be set to some viable size, say 100, for example.

2. In the second case, the trusting agent may know nothing of the other agent or the situation, in this case:

$$\text{Perceived\_Competence}_x(y, \alpha_x) = T_x \tag{3}$$

Note that it is possible for the trusting agent to know of the agent to be trusted, but not within the particular situation they find themselves in. It follows that the agent may have some form of trust in the other already. In this case, instead of T_x, we could use the value of $T_x(y)$ as x's estimate of the competence of y in the situation. i.e.:

$$\text{Perceived\_Competence}_x(y, \alpha_x) = T_x(y) \tag{4}$$

This is used in the example below.

Perceived Risk. The Perceived_Risk for x is simpler. Risk involves a weighing up of the costs and benefits of situations — whether it is worth risking the costs in order to obtain the benefits of the situation being resolved. It is also a subjective measure, for the most part. A risk that is acceptable for one may be too high for another [42]. The more information regarding the situation the agent has, the more able it is to estimate the risks involved. Zeckhauser, 1990, suggests that "Information is valuable when it accurately represents the risks posed. For one-time only decisions, from the standpoint of Bayesian decision making, the mean assessment of the probability of each outcome is all that matters, for that gives the likelihood with which the outcome will be received." [42, page 562]. For the purpose of the example below, a sample figure will be substituted here. A weighing up of the costs and benefits would give us the figure for utility, which may well be sufficient. We suggest the following, with certain reservations, for agent x with a sample, A, of situations experienced in the past:

$$\text{Perceived\_Risk}_x(\alpha_x) = \frac{1}{|A|} \sum_{\alpha \in A} \frac{C_x(\alpha_x)}{B_x(\alpha_x)} \tag{5}$$

The cooperation threshold is higher the more the importance of the situation. Hence the more important the situation, the more trust is necessary to enter into a cooperative situation with another agent. In addition, if the agent is trusted a lot, the cooperation threshold will be lower than for an agent who is trusted very little. Again, we use the estimate of trust given above. Here, however, it signifies the considerations the agent makes with respect to how well situations have turned out in the past with this particular trustee. If degrees of cooperation are possible, then this threshold could be *stepped* to take these degrees into account, and cooperation of a limited kind can be entered into with agents who are not trusted enough to cooperate with to the full extent required in the situation. Using the concept of trust alone is a limited approach to decisions such as whether or not to work with another agent. The idea of the importance of the situation takes this into account. Coupled with trust, this presents a powerful tool for an agent in decision making. Note that there may be situations where an agent has no choice but to trust another, for an example, see below. In this case, although the cooperation threshold may be high, cooperation will still occur. However, the agent can recognise the problems inherent in this, and make allowances, in the form of *safety nets*, in case of untrustworthy behaviour.

7.3 An Example

The formulae above are heuristic in nature, and are not intended to represent the final workings of trust, in human or machine agents. They do, however, provide an agent with a useful tool for the evaluation of situations and potential situations and cooperative relationships. In order to illustrate this further, this section provides an example of trust in action using the above formulae. The situation is adapted from Connah and Wavish, [10], and represents a problem involving furniture removal. In the situation, there are three agents and two pieces of furniture. Two of the agents, x and y, have the goal of moving a piece of furniture (different from the other agent's) to the door. Unfortunately, they cannot lift the pieces by themselves, and thus need help. Agent z is a professional furniture moving agent, whose services are for hire. Both x and y can consider who to ask for help, and who they will, in turn, help. For more information and an introduction to the problem, see Connah and Wavish [10].

Although an example using two agents is informative, it presents problems, not least the problem of deadlock should neither agent trust the other. With three agents or more, the concept of trust becomes more interesting, where different agents can be considered in the light of experience, competence, and so forth. Trust is, after all, a social phenomenon [26].

We will consider the situation from x's point of view. An interesting twist to the situation will be introduced in the fact that agent z is known and little trusted in general by x, but x 'knows' that z is, by chance, a specialised furniture removal agent, and thus extremely competent in this situation. Table 2 shows the 'thoughts' of agent x on cooperating with y or z.

The results of the deliberations are that x will favour working with y in this situation, despite z being a professional furniture mover. It is hoped that this mirrors real life situations to some extent — we may trust our best friend to help us move furniture rather than a professional, who may, perhaps, have less respect for our items of furniture.

8 Incomparability

Since we have introduced values to the notion of trust, it may be necessary to say a few words about how those values compare to each other. In other words, how is it possible to order the different trust values, and how does trust 'distribute' through agents.

It is not possible to say that x trusts z more than y, even if $T_x(z) = 0.8$ and $T_y(z) = 0.7$. There is no order there. This is because trust is such a subjective (i.e. agent-referenced) notion. What x calls 0.8, y may call 0.5.

x's value for	For y ($W = y$)	For z ($W = z$)
$\widehat{T_x(W)}$	0.8	0.3
$U_x(\alpha_x)$	0.85	0.85
$I_x(\alpha_x)$	0.75	0.75
$T_x(W, \alpha_x)$	0.51	0.19
$\text{Risk}_x(\alpha_x)$	0.75	0.75
$\text{Competence}_x(W, \alpha_x)$	0.8	0.9
$\text{Coop\_Thresh}_x(W, \alpha_x)$	0.35	0.47
Decision	Cooperate	Not cooperate

Table 2. Agent x's 'thoughts' on trusting agent y or agent z.

What implications does this have? It suggests firstly that trust is not a transitive relationship. From the point of view of an agent who knows some agents and not others, this means that estimating original trust in agents unknown to himself is difficult. It was originally envisaged that such an estimate could be based precisely on the views of other agents, if there were any.[6] This is still the case, but more thought is necessary on the part of the trusting agent. It is suggested that, for the present, the following rule should apply (assume the situation where x knows and trusts y, but not z, and y knows and has a trust value for z):

$$T_x(z) = T_y(z) \times T_x(y) \tag{6}$$

Note that, if one or both of these values are negative, the result should be set negative, no matter what sign it has.

9 The Limitations of Trust

The concept of trust introduced in this paper has its limitations. Not least is that there is a considerable amount of work still to be done. For example, thus far we have considered only one of the many different areas where trust can prove useful, namely that of initiating, or considering, cooperation with other agents. Whilst this is useful, it is only one area. Other important areas were briefly discussed above, notably the use of trust for gauging confidence in the information given to an agent, as in source control [16].

In addition, trust can help in the formation of groups: "Perhaps there is no single variable which so thoroughly influences interpersonal and group behaviour as does trust..." [18, page 131]. The formation of a group is basically an extended cooperative relationship between more than two agents. As such, the same kind of trust may well be used in forming the group. A question may arise in how groups themselves trust. For example, does a group trust another group more

[6] One of the benefits of a formalism using values is that such things are possible.

or less than another individual. Does a group see itself as a single entity, or just many individuals working to a common purpose? All these questions have yet to be addressed.

Trust provides an agent with a tool for judging the future, based on experience of the past. It is, however, limited when applied alone. It is not envisaged that an agent use trust as the sole measure of certainty in situations, much less as the only means of making a decision. The examples given above are intended to illustrate how trust can be used in such situations, but they are limited, as might be expected. When coupled with other methods, such as utility theory or theories of rational behaviour (e.g. [34]), trust provides an agent with a powerful and useful tool when interacting with other agents.

The heuristics given above present their own considerations. One of the problems with attaching values to things of that nature appears to be the different interpretations that can come from the values. It is hoped that the implementation of a trusting agent will enable a thorough investigation of these heuristics. They are, as such, introductory, but solid in that they work in a way in which trust would be expected to act in an agent, and allow for experimentation and discussion.

There are also problems with deadlock. Consider an agent in the furniture moving example who need help, but the only other agent around refuses to give it, and needs no help for itself (no furniture to move?). Nothing will happen — a classic deadlock situation, it seems. This suggests that such situations need reciprocity. This is, however, not so much a problem with trust as with cooperative situations in general. Nevertheless, it is a problem which will need addressing.

Finally, there is a large amount of complexity involved for one agent when considering many other in terms of trust. This may or may not be unavoidable, but it does suggest that trust may not be of much help in situated agents, where the behaviour is dictated by the environment. It will, however, help an agent when considering the environment and possible environment states, as briefly mentioned above.

10 Other Work

Trust is of course a wide ranging subject. We briefly raise a few interesting points, without attempting to be encyclopedic:

Numaoka [32], presents a view of what he calls the 'Special Interest Group', or SIG, which he states has "the most fundamental style of every organisation." A basic proposal in his paper is the idea of agents within a group voting to let another agent join the group. This could be extended using our definition of trust proposed here, such that the vote can be influenced, if not determined, by the trust the members of the group have in the prospective new member. This, in fact, provides an insight to the theory proposed here, in terms of group trust, which could be seen as the amount of trust a group has in an individual, whether inside or outside of the group, and also in another group (again, inside or outside of the group in question). Indeed, social choice is an aspect of decision making that may well benefit from a formal definition of trust.

Jennings [21] provides us with a view of "Joint Responsibility", in terms of what an agent states it will do, and the responsibility it has to the group as a whole. This, if developed further, could be used to influence the trust factor of our agents. For example, if an agent does not carry out the tasks it says it will, and is thus irresponsible, then we can use this irresponsibility as some form of metric towards altering our trust values accordingly.

There is some concern over the problem of security in distributed systems, in that "...the notion of trust in distributed systems is poorly understood. A satisfactory formal explication of trust has yet to be proposed."[41, page 40]. The idea of trust provided in this paper may go some way towards helping in that area.

Finally, in CSCW, we envisage that trust between humans and computers will become increasingly important, and a knowledge of its workings more useful over time. A recent paper, which develops the current work, presents this view [39]. Further work is presently ongoing in this area.

11 Conclusions and Ongoing Work

The notion of trust presented here is extendable. An implementation is in progress using HyperCard on the Macintosh which will show some of the usefulness of trust for artificial agents when considering cooperative situations. Further work will concern the decisions an agent makes with regard to its environment; how the agent trusts what it interacts with in a physical sense. Secondly, the fact that an agent trusts another does not impel the trusted agent to behave in a 'proper', trustworthy manner, although the fact that the other agent is trusting implies an acceptance of this 'danger'. What would be useful would be to include in the trusting world a law of some sort, not unlike the law-governed systems of Minsky [31, 30]. These systems provide laws which "...affect only what an agent may be able to do, not what it actually will do." [31, page 291]. What this means to the trusting agent is that the agent being trusted may be malevolent, but not capable of being malicious, and of actually damaging the trusting agent. In such a situation, trust may be built up more easily and safely, and this will be further studied in the future. In addition to the work discussed in section 9, further work will investigate the possibilities of using the concept of trust in, for example, distributed operating systems, as an analysis tool for more complex multi-system interactions, and so forth. Finally, the use of trust as a tool for the reduction of complexity, as briefly mentioned above, is an interesting avenue of work in itself.

A concept of trust has been introduced for agent interactions. The benefits of the concept include the allowance of a dynamic reference to interactions between agents, both with reference to the agent itself and to interactions with other agents. It allows at least minimal interactions and cooperation between agents where ordinarily there would be none. Most importantly the agents concerned are making an implicit acknowledgement of the possibility of malevolence or mistakes on behalf of the other agents, and as such there can be some form of backup against such occurences. It is thus of benefit both in that interactions are allowed, but with an acknowledgement of the problems involved. The formalism allows us to clarify our conception of the notion of trust, and, because of its inherent simplicity, allows a logical stepwise refinement of our understanding.

Acknowledgements

Many thanks to Harold Thimbleby, Lynne Coventry, Andy Cockburn, Steve Jones, and others for constructive comments on earlier drafts of this paper. Also to the attendees of the MAAMAW'92 workshop in Italy for instructive insights into some of the aspects of trust. The comments from anonymous referees enabled the author to improve this paper a great deal, for which, many thanks. The major part of this work was carried out when the author was a SERC CASE student, partially sponsored by Canon Research Europe.

References

1. Robert Axelrod. The evolution of strategies in the prisoner's dilemma. In Lawrence Davis, editor, *Genetic Algorithms and Simulated Annealing*, pages 32–41. Pitman, London, 1987.
2. Robert Axelrod. *The Evolution of Cooperation*. Penguin Books, London, 1990.
3. Annette Baier. Trust and antitrust. *Ethics*, 96(2):231–260, January 1986.
4. Bernard Barber. *Logic and Limits of Trust*. Rutgers University Press, New Jersey, 1983.
5. Roy L. Behr. Nice guys finish last - sometimes. *Journal of Conflict Resolution*, 25(2):289–300, June 1981.
6. Richard Boyle and Phillip Bonacich. The development of trust and mistrust in mixed-motive games. *Sociometry*, 33:123–139, 1970.
7. Alexander Broadie. *Trust*. Presentation given for the Henry Duncan prize, the Royal Society of Edinburgh, 2nd December, 1991.
8. Stephanie Cammarata, David McArthur, and Randall Steeb. Strategies of cooperation in distributed problem solving. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1983.
9. Man Kit Chang and Carson C. Woo. SANP: A communication level protocol for negotiations. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany*, 1991.
10. David Connah and Peter Wavish. An experiment in cooperation. In Yves Demazeau and Jean-Pierre Muller, editors, *Decentralized AI*, pages 197–212. Elsevier Science Publishers (North-Holland), 1990.
11. Partha Dasgupta. Trust as a commodity. In Diego Gambetta, editor, *Trust*, chapter 4, pages 49–72. Blackwell, 1990.
12. Morton Deutsch. Cooperation and trust: Some theoretical notes. In M. R. Jones, editor, *Nebraska Symposium on Motivation*. Nebraska University Press, 1962.
13. Morton Deutsch. *The Resolution of Conflict*. Yale University Press, New Haven and London, 1973.

14. Julia Rose Galliers. A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict. Technical Report No. 172, University of Cambridge Computer Laboratory, 1989.
15. Diego Gambetta, editor. *Trust*. Basil Blackwell, Oxford, 1990.
16. Roberto Garigliano, Albert Bokma, and Derek Long. A model for learning by source control. In Bouchon, Saiger, and Yager, editors, *Uncertainty and Intelligent Systems*, pages 163–170. Springer Verlag, Lecture Notes in Computer Science, LNC 313, 1988.
17. Matthew L. Ginsberg. Decision procedures. In M. Huhns, editor, *Distributed Artificial Intelligence, Volume 1*, chapter 1, pages 3–28. Pitman, London, 1987.
18. Robert T. Golembiewski and Mark McConkie. The centrality of interpersonal trust in group processes. In Cary L. Cooper, editor, *Theories of Group Processes*, chapter 7, pages 131–185. Wiley, 1975.
19. Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, July 1990.
20. Lars Hertzberg. On the attitude of trust. *Inquiry*, 31(3):307–322, September 1988.
21. Nick R. Jennings. On being responsible. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany, Panel Session*, 1991.
22. John T. Kohl. The evolution of the Kerberos authentication service. In *Europen '91 - European Conference on Open Systems*, pages 295–313, 1991.
23. Olli Lagenspetz. Legitimacy and trust. *Philosophical Investigations*, 15(1):1–21, January 1992.
24. Bernhardt Lieberman. *i*-Trust: a notion of trust in three-person games and international affairs. *Journal of Conflict Resolution*, 8(3):271–280, 1964.
25. Niklas Luhmann. *Trust and Power*. Wiley, Chichester, 1979.
26. Niklas Luhmann. Familiarity, confidence, trust: Problems and alternatives. In Diego Gambetta, editor, *Trust*, chapter 6, pages 94–107. Blackwell, 1990.
27. Stephen Marsh. Optimism, pessimism, and trust (working title). Department of Computing Science, University of Stirling, 1993.
28. Stephen Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Computing Science, University of Stirling, In preparation, 1994.
29. Steven McNeel. Training cooperation in the prisoner's dilemma. *Journal of Experimental Social Psychology*, 9:335–348, 1973.
30. Naftaly H. Minsky. The imposition of protocols over open distributed systems. *IEEE Trans. Software Engineering*, pages 183–195, February 1991.
31. Naftaly H. Minsky. Law-governed systems. *Software Engineering Journal*, pages 285–302, September 1991.
32. Chisato Numaoka. Conversation for organisational models. In *Pre-Proceedings MAAMAW'91: Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Germany, Panel Session*, 1991.
33. Derek Parfit. Prudence, morality, and the prisoner's dilemma. In Jon Elster, editor, *Rational Choice*, pages 34–59, 1986.
34. Jeffrey S. Rosenschein. *Rational Interaction: Cooperation among Intelligent Agents*. PhD thesis, Stanford University, October, 1985.
35. John F. Shoch and Jon A. Hupp. The "worm" programs: early experience with a distributed computation. *Communications of the ACM*, 25(3):172–180, March 1982.
36. Eugene H. Spafford. The Internet Worm: Crisis and Aftermath. *Comms. ACM*, 32(6):678–687, June 1989.

37. Robert L. Swinth. The establishment of the trust relationship. *Journal of Conflict Resolution*, 11(3):335–344, 1967.
38. Harold Thimbleby. Can viruses ever be useful? *Computers and Security*, 10:111–114, 1991.
39. Harold Thimbleby, Steve Marsh, Steve Jones, and Andy Cockburn. Trust in CSCW. In Steve Scrivener, editor, *Computer Supported Cooperative Work*. Ashgate Publishing, 1993.
40. Donnell Wallace and Paul Rothaus. Communication, Group Loyalty, and Trust in the PD Game. *Journal of Conflict Resolution*, 13(3):370–380, 1969.
41. Thomas Y. C. Woo and Simon S. Lam. Authentication for distributed systems. *IEEE Computer*, pages 39–52, January 1992.
42. Richard J. Zeckhauser and W. Kip Viscusi. Risk within reason. *Science*, 248:559–564, May 4th 1990.

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].

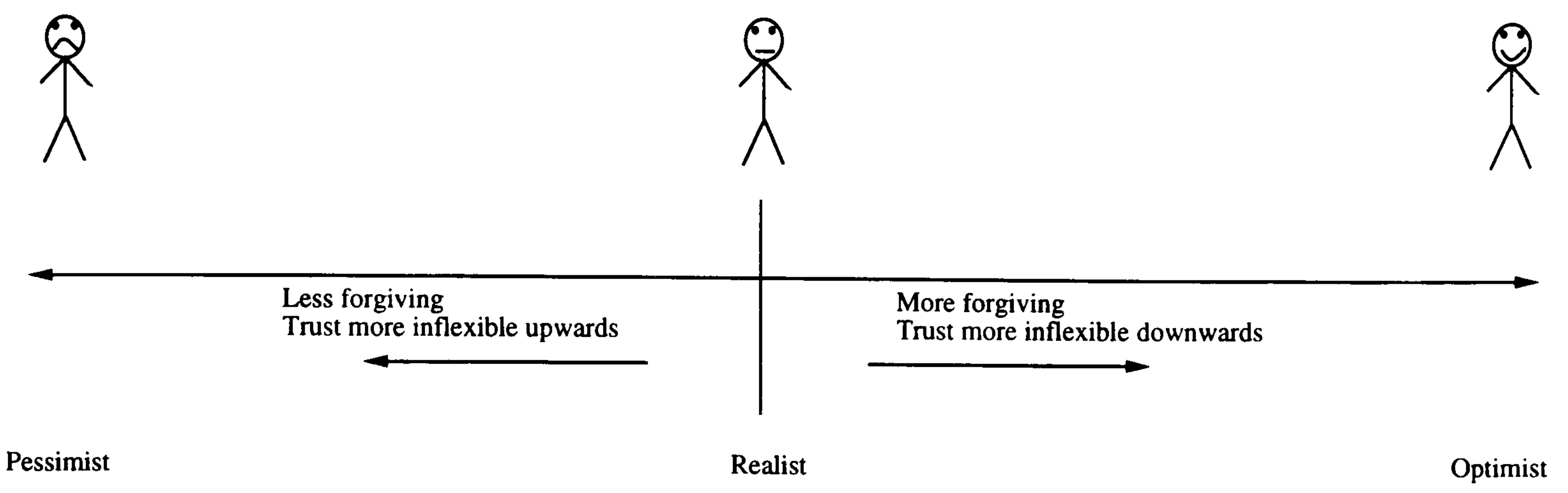

Figure 1: Spectrum of Realism

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 ψ_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 ϵ_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.

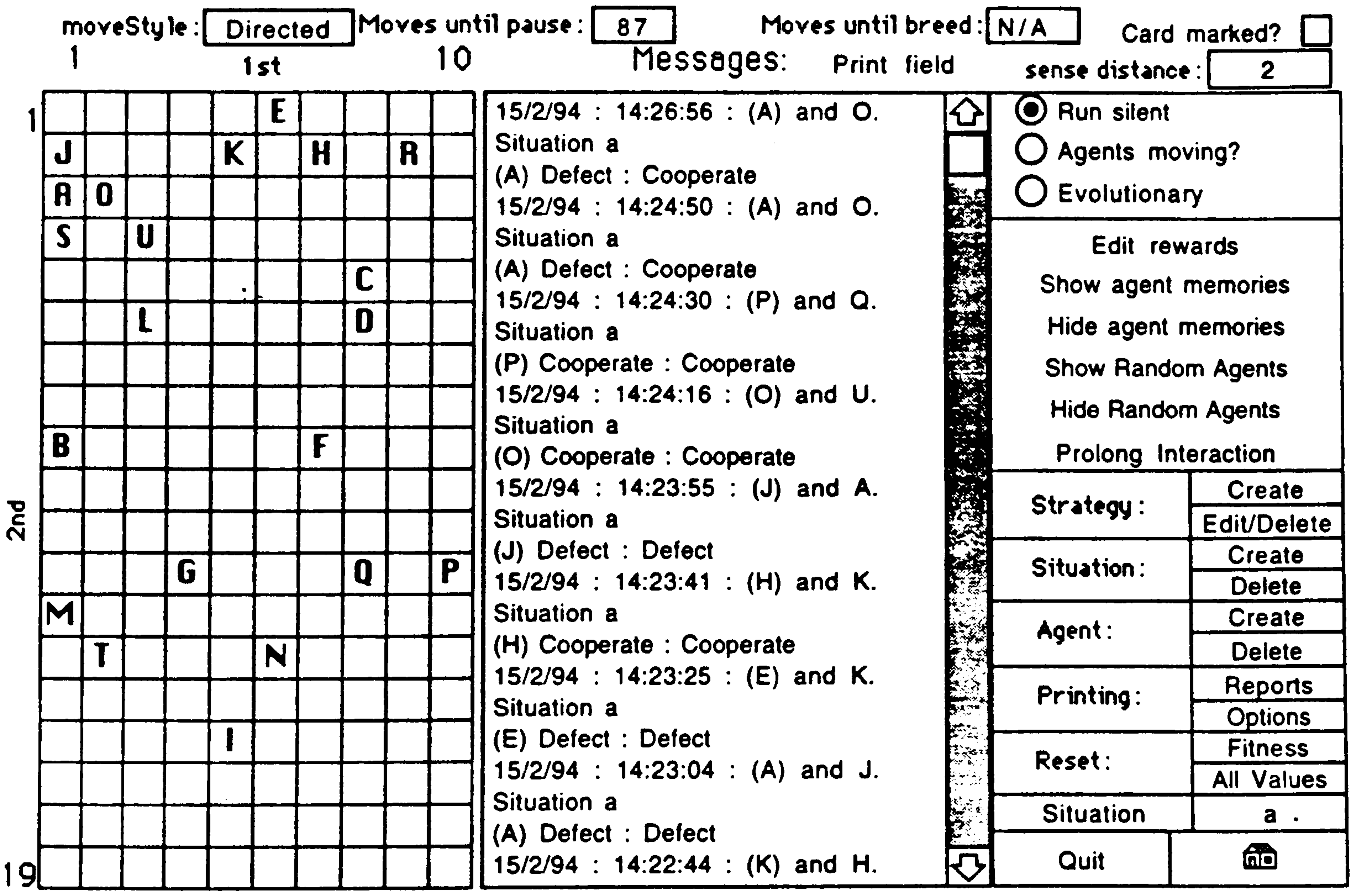

Figure 2: A simple testbed for determining trusting agents' behaviour..

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.

For Johnstone,
who lived fast and trusted much.