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 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.


3 Trust vis a vis DAI

Implicit in the notion of DAI is the concept of decentralisation. Since decentral-
isation 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 other-
wise be fairly damaging — a form of 'safety net'.

¹ This, again, is where the formalism can help, in clarifying the situation.