Formal Analysis of Trust Dynamics in Human and Software Agent Experiments

Tibor Bosse[1], Catholijn M. Jonker[2], Jan Treur[1], and Dmytro Tykhonov[2]

[1] Department of Artificial Intelligence, Vrije Universiteit of Amsterdam, 1081HV, Amsterdam, The Netherlands
{tbosse,treur}@few.vu.nl
[2] Man-Machine Interaction Group, Delft University of Technology, Mekelweg 4, Delft, The Netherlands
{catholijn,dmytro}@mmi.tudelft.nl

Abstract. Recognizing that trust states are mental states, this paper presents a formal analysis of the dynamics of trust in terms of the functional roles and representation relations for trust states. This formal analysis is done both in a logical framework and in a mathematical framework based on integral and differential equations. Furthermore, the paper presents formal specifications of a number of relevant dynamic properties of trust. The specifications provided were used to perform automated formal analysis of empirical and simulated data from two case studies, one involving two experiments with humans, and one involving simulation experiments in the context of an economic game.

Keywords: Trust, dynamics, intelligent agents, human experiments.

1 Introduction

In the literature, a variety of definitions can be found for the notion of trust. The common factor in these definitions is that trust is a complex issue relating to belief in honesty, faithfulness, competence, and reliability of the trusted system actors, see e.g., [5, 6, 7, 13, 14, 19, 22, 23, 27]. Furthermore, the definitions indicate that trust depends on the context in which interaction occurs or on the observer's point of view. The agent might for example trust that the statements made by another agent are true. Likewise, the agent might trust the commitment of another agent with respect to certain (joint) goals, or the agent might trust that another agent is capable of performing certain tasks.

In [21] trust is analysed referring to observations which in turn lead to expectations: 'observations that indicate that members of a system act according to and are secure in the expected futures constituted by the presence of each other for their symbolic representations.' In [8] it is agreed that observations are important for trust, and trust is defined as: 'trust is the outcome of observations leading to the belief that the actions of another may be relied upon, without explicit guarantee, to achieve a goal in a risky situation.' Elofson [8] notes that trust can be developed over time as the outcome of a series of confirming observations. The evolution of trust over time, also called the dynamics of trust, is addressed in this paper.

We conceive trust as an internal (mental) state property of an agent that may help him to conduct various kinds of complex behaviour. The cognitive concept trust

M. Klusch et al. (Eds.): CIA 2007, LNAI 4676, pp. 343–359, 2007.

enables the agent to effectively cope with complex environments that are populated by self-interested agents. Trust is based on a number of factors, an important one being the agent's own experiences with the subject of trust; e.g., another agent. Each event that can influence the degree of trust is interpreted by the agent to be either a trust-negative experience or a trust-positive experience. If the event is interpreted to be a trust-negative experience, the agent will lose its trust to some degree, if it is interpreted to be trust-positive, the agent will gain trust to some degree. The degree to which trust is changed depends on the characteristics of the agent. Agents equipped with a concept of trust perform a form of continual verification and validation of the subjects of trust over time. For example, a car is trusted, based on a multitude of experiences with that specific car, and with other cars in general.

In this paper the dependence of trust on experiences obtained over time is the main focus, abstracting from other possible influences, which (as a form of idealisation) are assumed not present or constant. The dynamics of trust are formalised by sequences of mental states. The functional roles and the representation relations of trust states are formalised both in a logical and in a numerical mathematical way based on integral and differential equations. Properties of trust dynamics are formalised accordingly, incorporating logical and numerical aspects. Empirical and simulated data from two experiments with trust are formally analysed using an automated checker for such properties against traces [2]. The first experiment, called the scenario case study, studies dynamics of trust of humans when confronted with two different scenario's. The other experiment, called the Trust & Tracing case study, is an agent-based simulation of an economic game.

Section 2 sketches the preliminaries for modelling trust dynamics using both mathematical and logical means. To properly catch the dynamics of trust as a cognitive phenomenon, Section 3 describes trust states as mental states for which the functional or causal roles and representation relations can be analysed and formally specified in logical terms. In Section 4 a continuous approach to trust dynamics is presented showing how functional role and representation relation specifications for trust states can be defined within Dynamical Systems Theory [25]. Section 5 presents a number of formally specified properties of trust used to analyse the case studies. The scenario case study is presented in Section 6; its formal analysis, based on properties formalised in Section 5, is presented in Section 7. The Trust & Tracing case study is presented in Section 8 and analysed in Section 9, using the properties from Section 5. Section 10 is a discussion.

2 Preliminaries

In this paper, trust is considered a mental agent concept that depends on experiences. This is modelled by a mathematical function that relates sequences of experiences to trust representations: a *trust evolution* function. As a computational alternative the iterative *trust update* function relates a current trust representation and a current experience to the next trust representation. To obtain a formal mathematical framework, the following four sets are introduced. A partially ordered set EV of *experience values*, a linearly ordered time frame TIME with initial time point 0, the set ES of *experience sequences*, i.e., functions from TIME to EV, and a partially ordered set TV of *trust values*. Examples of such sets EV and TV are the (closed) interval of real

numbers between -1 and 1, the set {-1, 0, 1}, or sets of qualitative labels, such as {very high, high, neutral, low, very low}. Within the sets EV and TV, subsets of positive and negative values are distinguished. A *trust evolution function* is a function te : ES x TIME → TV. A *trust update function* is a function tu : E x TV → TV. Throughout this paper we assume that the value of a trust evolution function te at time t only depends on the experiences in the past of t (*future independence*).

A logical formalisation is based on the language TTL [2]; this is a language in the class of reified predicate logic-based temporal languages as distinguished in [11, 12]. In TTL, *state atoms* are atoms over the state ontology Ont, such as experience(v) and trust(w), which express the experience value v from sort EV and trust value w from sort TV in a state, and for relations for these sorts such as <, pos and neg. A *state* is a truth assignment (of truth values true and false) to the set of (ground) state atoms; STATES(Ont) denotes the set of all states over state ontology Ont. A *trace* for state ontology Ont is a function γ: TIME → STATES(Ont); they form the set TRACES(Ont). The expression state(γ, t) |= p denotes that state property holds in the state of γ at time t. Here |= is an infix predicate of the language. Based on such atoms formulae can be formed using the predicate logic connectives, including quantifiers (e.g., over time and traces).

3 Trust States and Mental States

Following literature on Philosophy of Mind, such as [20], a mental state can be characterised in two manners:

- by the functional role it plays, defining the immediate predecessor and successor states in causal chains in which it is involved
- by its representation relations, defining how the mental state relates to other states more distant in locality and time

For each of these two aspects, trust states will be analysed by a simple example.

The Shop Example

Consider the following example, concerning agent A and a specific shop. The behaviour of agent A considered is as follows:

- agent A can go to the shop or avoid it
- when meeting somebody, agent A can tell that it is a bad shop or that it is a good shop

The following types of events determine the behaviour of agent A

negative events:

- an experience that a product bought in this shop was of bad quality
- somebody else tells A that it is a bad shop
- passing the shop, A observes that there are no customers in the shop

positive events:

- an experience that a product bought in this shop was of good quality
- somebody else tells A that it is a good shop
- passing the shop, A observes that there are customers in the shop

Assume for the sake of simplicity that only the last two experiences count for the behaviour of A, and that the past pattern a considered are histories in which the last two experiences are negative events. The future pattern b considered are the futures in which the agent avoids the shop and, when meeting somebody tells that it is a bad shop. It is assumed that, viewed from an external perspective, past pattern a leads to future pattern b.

Functional Role Specifications for a Trust State

Functional roles are described from a backward perspective (relating the trust state to states that lead to it) and a forward perspective (relating the trust state to states to which it leads). The trust state property 'very negative' is used as an illustration. For the forward perspective a relatively simple specification can be made; for example:

Functional role specification: forward

If the agent has very negative trust about the shop, then it will avoid the shop and when meeting somebody, (s)he will speak negatively about the shop.

$$\text{state}(\gamma, t) \models \text{trust}(\text{very\_negative}) \Rightarrow \exists t1 \geq t \ [\ t \leq t1 \leq t+d \ \&\ \forall t2 \ [\ t1 \leq t2 \leq t1+d \Rightarrow$$
$$[\text{state}(\gamma, t2) \models \text{preparation\_for}(\text{avoiding\_shop}) \ \&$$
$$\text{state}(\gamma, t2) \models \text{conditional\_preparation\_for}(\text{meets}(A, B), \text{speaks\_bad\_about\_shop\_to}(A, B))\ ]\ ]$$

Here preparation for an action leads to performing the action (unless it is blocked), and conditional preparation for an action leads to preparation for the action as soon as the condition is observed. The backward perspective is inherently less simple, since trust is a type of mental state property that accumulates over longer time periods. This means that trust at a next point in time depends on present experiences but also on the present trust state. This gives it a recursive character. For example,

Functional role specification: backward

If the agent has negative or very negative trust about the shop, and it has a negative experience, then it will have very negative trust.

$$[\text{state}(\gamma, t) \models \text{trust}(\text{negative}) \vee \text{trust}(\text{very\_negative}) \ \&\ \text{state}(\gamma, t) \models \text{observes}(e) \ \&\ \text{neg}(e)\ ] \Rightarrow$$
$$\exists t1 \geq t \ [\ t \leq t1 \leq t+d \ \&\ \forall t2 \ [\ t1 \leq t2 \leq t1+d \Rightarrow \text{state}(\gamma, t1) \models \text{trust}(\text{very\_negative})\ ]$$

A numerical example with trust decay rate r is as follows:
If the agent has trust level w about the shop, and it has an experience of level v, then it will have trust of level rw+(1-r)v.

$$[\ \text{state}(\gamma, t) \models \text{trust}(w) \ \&\ \text{state}(\gamma, t) \models \text{observes}(\text{experience}(v)) \Rightarrow$$
$$\exists t1 \geq t \ [\ t \leq t1 \leq t+d \ \&\ \forall t2 \ [\ t1 \leq t2 \leq t1+d \Rightarrow \text{state}(\gamma, t1) \models \text{trust}(rw+(1-r)v)\ ]$$

Note that in mathematical terms a backward functional role specification corresponds to a trust update function, and can be used directly in a computational manner for simulation.

Representation Relations for a Trust State

Trust is an example of a mental state property that heavily relies on histories of experiences, as also is found in empirical work; e.g. [17]. By abstracting from these histories in the form of a trust state that accumulates the history of experiences, the future dynamics can be described on the basis of the present mental state in a simple manner. Here the past pattern can be characterised by a formula $\varphi(\gamma, t)$:

$$\exists t1 < t2 \leq t \ [\text{state}(\gamma, t1) \models \text{observes}(e1) \ \&\ \text{state}(\gamma, t2) \models \text{observes}(e2) \ \&$$
$$\text{neg}(e1) \ \&\ \text{neg}(e2) \ \&\forall t3 \ [\ t1 \leq t3 \leq t \Rightarrow \neg \exists e3 \ [\ \text{pos}(e3) \ \&\ \text{state}(\gamma, t3) \models \text{observes}(e3)\ ]\ ]$$

Moreover, the future pattern can be characterised by a $\psi(\gamma, t)$:

$\exists t1 \geq t$ state(γ, t1) |= performs(avoiding_shop) & $\forall$ t2 $\geq$ t [state(γ, t2) |= observes(meeting(A, B)) $\Rightarrow$
$\exists t3 \geq t2$ state(γ, t3) |= performs(speaking_bad_about_shop_to(A, B))]

This obtains the following temporal relational specifications for the representational content of the trust state

Representation relation: forward

$\forall \gamma, t$ [state(γ, t) |= trust(very_negative) $\Leftrightarrow$ $\psi(\gamma, t)$]

Representation relation: backward

$\forall \gamma, t$ [$\varphi(\gamma, t)$ $\Leftrightarrow$ state(γ, t) |= trust(very_negative)]

A numerical example with trust decay rate r involves summation over time as follows:

If t0 is a time point and d a duration and t1 = t0+d and the agent has at time point t0 trust level w(t0) and at time points t between t0 and t1 it has experiences of level v(t), then at t1 it will have trust of level $w(t1) = r^d w(t0) + (1-r)\sum_{0 \leq d1 \leq d-1} r^{d1} v(t0+d-d1)$.

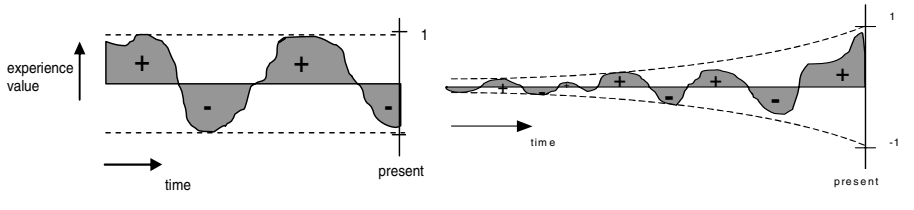
This property can be expressed in TTL as follows (here for any formula φ, the expression case(φ, v1, v2) indicates the value v1 if φ is true, and v2 otherwise):

[state(γ, t0) |= trust(w0) & $w1 = r^d w0 + (1-r) \Sigma_{k=0}^{d} \Sigma_{v:EV}$ case(state(γ, t0+k) |= experience(v), r^{d-k}v, 0)
$\Rightarrow$ state(γ, t0+d) |= trust(w1)]

In Section 4 an analysis of representation relations and functional role specifications for continuous trust values over continuous time in terms of the Dynamical Systems Theory [25] (based on integral and differential equations) will be presented.

4 A Continuous Approach

The notions functional role and representation relation can be analysed in a continuous form as well. This section considers a continuous mental state property for trust over continuous time. As an example, it is assumed that trust in the weather forecast depends on one's experiences based on continuously monitoring the actual weather and comparing the observed weather with the predicted weather. For some of the patterns of behaviour, decisions may depend on your trust in the weather forecast. In particular, the decision to take an umbrella depends not only on the weather forecast, but also on your trust in the weather forecast. For example, when the weather forecast is not bad, but



a b

Fig. 1. Trust for continuous experiences (a) without decay and (b) with decay

your trust is low, you still will take an umbrella with you. It is assumed that for each point in time your experience with the weather forecast is a modelled by value (real number) between -1 (negative experience) and 1 (positive experience).

As a first approach (for reasons of presentation leaving decay aside for a moment) the accumulation of experiences in trust may be described by averaging the accumulation of the shaded area of the graph of experiences values over time, shown in Figure 1. So, a trust state represents a kind of average of the experiences over time. More specifically, the trust value is taken to be the real number indicating the shaded area divided by the length of the time interval, where the parts below the time axis count as negative. Within an overall trace γ the relation between trust value $tv_\gamma(t)$ at a certain point in time $t > 0$ and the experience history can be modelled by the following backward representation relation for the trust state, expressed by an integral over time until t of the experience value $ev_\gamma(t)$, i.e.,

$$tv_\gamma(t) = \int_0^t ev_\gamma(u)\, du / t$$

where for a trace γ the functions tv, ev are defined by:

$tv_\gamma(t) = v$ iff state(γ, t, internal) |= has_value(trust, v)
$ev_\gamma(t) = w$ iff state(γ, t, input) |= has_value(experience, w)

This shows the cumulative character of a (backward) representation relation. A functional role specification has a more local character. Such a local relationship can be modelled by a differential equation, which can be found from the integral relation

$$t\ tv_\gamma(t) = \int_0^t ev_\gamma(u)\, du$$

by differentiation and application of the product rule as follows:

$$d/dt\ t.\ tv_\gamma(t) = d/dt \int_0^t ev_\gamma(u)\, du$$
$$(d/dt\ t)\ .\ tv_\gamma(t) + t.\ d/dt\ tv_\gamma(t) = ev_\gamma(t)$$
$$tv_\gamma(t) + t.\ d/dt\ tv_\gamma(t) = ev_\gamma(t)$$

Therefore the following differential equation is obtained:

$$dtv_\gamma(t) / dt = [\ ev_\gamma(t) - tv_\gamma(t)\ ] / t$$

In discretised form this provides:

$$tv_\gamma(t+ \Delta t) = tv_\gamma(t) + [[\ ev_\gamma(t) - tv_\gamma(t)\ ] / t\ ]\ \Delta t$$

This shows the backward functional role specification for a trust state, which has the form of a trust update function and can be used for simulation (based on Euler's method; also the more efficient higher-order Runge-Kutta methods are applicable).

The example shows that, for continuous models, characterisations of representation relations and functional role specifications show up that are formulated in terms of integrals or differential equations. This provides an interesting connection of the higher-level cognitive concept trust to the Dynamical Systems Theory as advocated, for example, in [25]. In the above example, it is not realistic that experiences very far back in time count the same as recent experiences. In the accumulation of trust, experiences further back in time will have to count less than more recent experiences, based on a kind of inflation rate, or increasing memory vagueness. Therefore a more realistic model is obtained if it is assumed that the graph of experience values against

time is modified to fit it between two curves that are closer to zero further back in time. Trust then is the accumulation of the areas in the graph below, where the parts below the time axis count negative.

The limiting curves can be based, for example, on an exponential function e^{at} with $a > 0$ a real number related to the strength of the decay. The graph in Figure 1b depicts the resulting function $ev_\gamma(t)\ e^{at}$. Under these assumptions the representation relation between trust and experiences can be modelled by the integral

$$tv_\gamma(t) = \int_0^t ev_\gamma(u)\ e^{au}\ du\ .\ a\ /\ (e^{at} - 1)$$

where the factor $a\ /\ (e^{at} - 1)$ is a normalisation factor to normalise trust in the interval [-1, 1]. As before, a functional role specification for the trust state in the form of the following differential equation can be obtained:

$$\begin{aligned} dtv_\gamma(t)/dt &= [\ ev_\gamma(t) - tv_\gamma(t)\ ]\ .\ a\ e^{at}\ /\ (e^{at} - 1) \\ &= [\ ev_\gamma(t) - tv_\gamma(t)\ ]\ .\ a\ \ /\ (1 - e^{-at}) \end{aligned}$$

5 Properties of Trust

In [18] a number of possible properties of trust evolution and trust update functions are defined in mathematical terms. For motivation and further explanation we refer to the reference mentioned. This paper contributes formalisation of the properties in logical terms. The following properties from [18] are considered (here $e,f \in ES$, $\gamma \in TRACES$, $s,t,u \in TIME$, $v \in EV$, $w \in TV$, and $e|\leq t$ denotes the restriction of sequence e to time points $\leq t$).

Future independence. Future independence expresses that trust only depends on past experiences, not on future experiences. This is a quite natural assumption that is assumed to hold for all trust evolution functions. In mathematical terms:

$$e|\leq t = f|\leq t\ \ \&\ te(e, 0) = te(f, 0) \quad \Rightarrow \quad te(e, t) = te(f, t)$$

This property can be expressed logically in TTL [2] as follows:

$$\begin{aligned} &\forall \gamma 1, \gamma 2, t\ \ [\ \forall w\ [\ state(\gamma 1, 0) \models trust(w) \Leftrightarrow state(\gamma 2, 0) \models trust(w)\ ]\ \& \\ &\forall t1 \leq t\ \ \forall v\ [\ state(\gamma 1, t1) \models experience(v) \Leftrightarrow state(\gamma 2, t1) \models experience(v)\ ]\ ] \Rightarrow \\ &\quad \forall w\ [\ state(\gamma 1, t) \models trust(w) \Leftrightarrow state(\gamma 2, t) \models trust(w)\ ] \end{aligned}$$

Note that as the property refers to two different histories (and compares them), it cannot be expressed in modal temporal logic: then only reference can be made to one history.

Limited memory d. Limited memory expresses that trust only depends on past experiences in a certain time interval of duration d back in time from the present. In mathematical terms:

$$e|\geq t-d = f|\geq t-d \ \Rightarrow \ \ te(e, t) = te(f, t)$$

This property can be expressed logically in TTL as follows:

$$\begin{aligned} &\forall \gamma 1, \gamma 2, t\ \ [\ \forall t1 \leq t\ \ [\ t1 \geq t-d \Rightarrow \\ &\quad \forall v\ [\ state(\gamma 1, t1) \models experience(v) \Leftrightarrow state(\gamma 2, t1) \models experience(v)\ ]\ ] \Rightarrow \\ &\quad\quad \forall w\ [\ state(\gamma 1, t) \models trust(w) \Leftrightarrow state(\gamma 2, t) \models trust(w)\ ] \end{aligned}$$

Trust Monotonicity. Monotonicity expresses that the more positive experiences are, the higher the trust. Mathematically:

$$e \leq f\ \ \&\ \ te(e, 0) \leq te(f, 0) \quad \Rightarrow \quad te(e, t) \leq te(f, t)$$

Note that this property again refers to two different histories; it can be expressed logically as follows:

$\forall\gamma1, \gamma2, t$
$[\ \forall w1,w2\ [\ state(\gamma1, 0) \models trust(w1)\ \&\ state(\gamma2, 0) \models trust(w2) \Rightarrow w1 \leq w2]\ \ \&\ \ \forall t1 \leq t\ \ \forall v1,v2$
$[\ state(\gamma1, t1) \models experience(v1)\ \&\ state(\gamma2, t1) \models experience(v2) \Rightarrow v1 \leq v2]\ \ ] \Rightarrow\ \ \forall w1,w2\ [\ state(\gamma1, t) \models trust(w1)\ \&\ \ state(\gamma2, t) \models trust(w2) \Rightarrow w1 \leq w2]$

The two different histories again imply that this property cannot be expressed in modal temporal logic.

Positive trust extension. Positive (or negative) trust extension expresses that trust is increasing if only positive (negative) experiences are encountered, i.e., after a positive (negative) experience, trust will become at least as high (low) as it was. In mathematical terms:

$\forall s,t\ \ [\forall u \in TIME\ [\ s \leq u < t\ \Rightarrow e_u\ \ positive\ \ ]\ \ \Rightarrow\ \ te(e, s) \leq te(e, t)$

This property can be expressed logically as follows

$\forall s,t\ \ [\forall u\ \forall v\ [\ s \leq u < t\ \ \&\ \ state(\gamma, u) \models experience(v)\ \ \Rightarrow\ \ pos(v)\ ]\ ]$
$\Rightarrow\ \forall w1,w2\ [\ state(\gamma, s) \models trust(w1)\ \&\ state(\gamma, t) \models trust(w2) \Rightarrow w1 \leq w2]$

Trust Flexibility: degree of trust gaining d. The property degree of trust gaining (or dropping) expresses, independent of the trust state, after how many positive (or negative) experiences trust will be positive (or negative). In mathematical terms:

$\forall t\ \ [\forall k \in TIME\ [\ t\text{-}d < k \leq t\ \Rightarrow e_k\ negative\ ]\ \Rightarrow\ \ te(e, t)\ negative$

This property can be expressed logically as follows

$\forall t, d\ \ [\ state(\gamma, t) \models trust(w)\ \&\ \forall t1\ \ [\ [t\text{-}d \leq t1 \leq t\ \ \&\ \ state(\gamma, t1) \models experience(v)\ ]\ \Rightarrow pos(v)\ ]\ \Rightarrow\ \ pos(w)\ ]$

Positive limit approximation. Positive limit approximation expresses that it is always possible to reach maximal trust, if a sufficiently long period with only positive experiences is encountered (the same for the negative case). Mathematically:

If a t exists such that for all $s > t$ it holds that e_s is maximal (in EV),
then a t' exists such that $te(e, t)$ is maximal (in TV) for all $t > t'$.

This property can be expressed logically as follows:

$\exists t\ \forall t1 \geq t\ [\ state(\gamma, t1) \models experience(v) \Rightarrow \neg\ \exists v1\ v1 > v\ ] \Rightarrow\ \ \exists t'\ \forall t1 \geq t'\ [\ state(\gamma, t1) \models trust(w) \Rightarrow \neg\ \exists w1\ w1 > w\ ]$

This property can be relaxed a bit by specifying a margin such that trust should become that close to the maximal value. In [18] also the following properties of trust update functions are defined and related to properties of trust evolution functions. They are formalised logically as follows.

Trust Update Monotonicity. A trust update function tu is monotonic if higher experience values and higher trust values lead to higher trust update values. In mathematical terms:

$ev1 \leq ev2\ \ \&\ \ tv1 \leq tv2\ \ \ \ \Rightarrow\ \ \ \ tu(ev1, tv1) \leq tu(ev2, tv2)$

This property can be expressed logically as follows

$\forall\gamma1, \gamma2, t\ \forall v1,v2, w1, w2\ [state(\gamma1, t) \models experience(v1) \wedge trust(w1)\&$
$state(\gamma2, t) \models experience(v2) \wedge trust(w2) \Rightarrow v1 \leq v2\ \&\ w1 \leq w2]\ \Rightarrow$
$\forall w1,w2\ [\ state(\gamma1, t+1) \models trust(w1)\ \&\ state(\gamma2, t+1) \models trust(w2) \Rightarrow w1 \leq w2]$

Positive and negative trust extension. This property states that positive (negative) experiences lead to higher (lower) trust values. In mathematical terms:

$ev\ \ positive\ \Rightarrow\ tu(ev, tv) \geq tv$

This property can be expressed in logical terms as follows:

$\forall v1,w1,w2\ [\ [\ state(\gamma, t) \models experience(v1) \wedge trust(w1)\ \&\ state(\gamma, t+1) \models trust(w2)\ ] \Rightarrow w1 \leq w2]$

Strict positive (negative) monotonic progression. This property is a stronger version of the previous one and states that positive (negative) experiences lead to strictly higher (lower) trust values, as long as this is possible. In mathematical terms:

ev positive and tv not maximal (in T) $\Rightarrow$ tu(ev, tv) > tv

This property can be expressed logically as follows

$\forall v1,w1,w2\ [\ [\ state(\gamma, t) \models experience(v1) \wedge trust(w1)\ \&\ \exists w3\ w3>w1\ \ \&\ \ state(\gamma, t+1) \models trust(w2)\ ] \Rightarrow w1<w2]$

Negative or positive trust fixation of degree d. After d negative events the agent will never trust anymore and its trust will remain the least possible. After d positive events the agent will forever trust (even when faced with negative events) and its trust will remain maximal.

$\forall t, d\ [\ state(\gamma, t) \models trust(w)\ \&\ \forall t1\ [\ [t-d \leq t1 \leq t\ \ \&\ \ state(\gamma, t1) \models experience(v)\ ] \Rightarrow pos(v)\ ] \Rightarrow$
$\forall t2 \geq t\ [\ state(\gamma, t) \models trust(w2) \Rightarrow \neg\exists w\ \ w>w2\ ]$

6 The Scenario Case Study

This section describes an experiment based on 294 subjects, taken from [17]. From the 294 subjects, 238 subjects (81%) completed the full questionnaire. The other 19% of the subjects were either not able to complete the questionnaire because of technical problems, decided to stop during the experiment, or did not respond to a question within a given time limit of 15 minutes between each two questions. Only the data obtained from subjects that fully completed the questionnaire have been used. In the

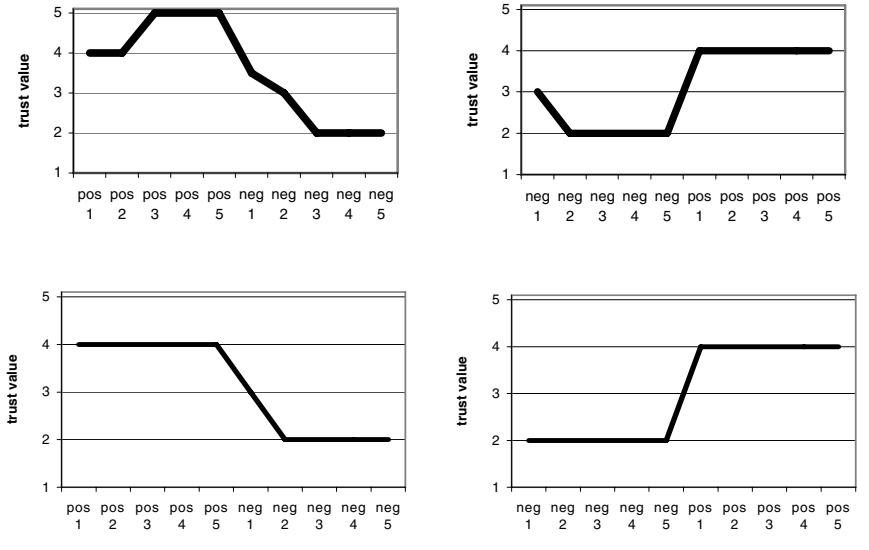


Fig. 2. Dynamics of trust in Photocopier (upper two), resp. Travel Agency (lower two) experiments, both for positive experiences first and negative experiences first

test the effect of experiences with an organisation or an object on the trust in that organisation or object is measured. The effect of the experiences on trust is measured by describing various experiences in small stories and instructing the subject to express his or her trust in the object or organisation after having gone through such experiences, using a five-points trust rating scale [17]. In this scale, trust value 3 represents neutral trust, value 4 and 5 represent positive trust, and value 1 and 2 represent negative trust.

Each scenario consisted of an introduction and ten distinctive stories, five of which were positive (written and validated as to induce trust) and five of which were negative (written and validated as to induce distrust). For more details, see [17]. In Figure 2 the median of the trust values of the population for both the photocopier and the travel agency are plotted. It shows that indeed trust increases when the subject had a positive experience and that trust decreases upon a negative experience. There is a clear difference between plots starting with negative experiences and plots that start with positive experiences. To determine the significance of this difference, a 2-way between subjects ANOVA test was performed on the means of the positive and negative experiences within a single scenario. The ANOVA test takes both the experience (positive or negative) and the order in which experiences are presented (positive-first or negative-first) into account. The results show that both factors have an effect on trust in the object or organisation at a significance level beyond 0,001.

7 Analysis of the Scenario Case Study

The outcomes of the experiments as described in Section 6 and (in particular) the traces shown in Figure 2 have been compared to the most relevant dynamic properties from Section 5:

Positive and negative trust extension. These properties are satisfied for all four types of human traces shown in Figure 2.

Strict positive and negative monotonic progression. These properties failed according to the TTL checker [2] for all four human traces, probably due to the low number of possible trust values (if there are few trust values to choose from, people do not always choose a new trust value after a new experience).

Trust Flexibility. The property succeeds for all four human traces, given the right value for d. For example, for the photocopier, 3 negative experiences in a row are sufficient to get a negative trust (no matter how positive trust was), and for the travelling agency 2 negative experiences are sufficient. For the positive side, in both cases only 1 positive experience is sufficient to get trust positive again, so the property 'degree of trust gaining 1' holds for both cases. An effect that does occur, however, in the photocopier context, is that after a series of negative experiences, the level of trust does not become as high as in the case of no negative experiences (see Figure 2). More refined properties than the ones above can be formulated to account for this relative form of trust fixation. Notice that in the travelling agency context this effect does not occur.

Positive limit approximation. This property assumes that the traces under investigation are of infinite length. Therefore, it makes no sense to check it against the traces of this case study. (For example, if for t one chooses the last time point of the trace, then the property is not very informative). However, analytically it can be shown that for the photocopier an upper limit of 5 and a lower limit of 2 are reached, whilst for the travel agency an upper limit of 4 and a lower limit of 2 are reached.

Limited memory d. According to the TTL checker, this property succeeds for all four human traces, given the right value for d. This can be illustrated by considering the same example as for Trust Flexibility above: for the photocopier, 3 negative experiences in a row are sufficient to get a negative trust (no matter what the previous trust value was), and for the travelling agency 2 negative experiences are sufficient.

8 Trust and Tracing Case Study

The Trust and Tracing game [24] is a research tool designed to study human behaviour with respect to trust in commodity supply chains and networks in different institutional and cultural settings. The game played by human participants is used both as a tool for data gathering and as a tool to make participants feed back on their daily experiences.

The focus of study is on trust in a business partner when acquiring or selling commodities with invisible quality. There are five roles: traders (producers, middlemen and retailers), consumers and a tracing agency. The real quality of a commodity is known by producers only. Sellers may deceive buyers with respect to quality, to gain profits. Buyers have either to rely on information provided by sellers (Trust) or to request a formal quality assessment at the Tracing Agency (Trace). This costs a tracing fee for the buyer if the product is what the seller stated (honest). The agency will punish untruthful sellers by a fine. Middleman and Retailers have an added value for the network by their ability to trace a product cheaper than a consumer can.

Commodities usually flow from producers to middlemen, from middlemen to retailers and from retailers to consumers. Players receive 'monopoly' money upfront. Producers receive sealed envelopes representing lots of commodities. Each lot is of a certain commodity type (represented by the colour of the envelope) and of either low or high quality (represented by a ticket covered in the envelope). The envelopes may only be opened by the tracing agency, or at the end of the game to count points collected by the consumers. The player who has collected most points is the winner in the consumer category. In the other categories the player with maximal profit wins.

Sessions played until 2005, see [16], provided insights, such as: participants who know and trust each other beforehand tend to start trading faster and trace less. The afterwards indignation about deceits that had not been found out during the game is higher in these groups than it is when participants do not know each other. The objective of [16] was to model the behaviour of sellers and buyers using the concept of trust. The model is inspired by [19] who define reliability trust as "trusting party's probability estimate of success of the transaction". This choice allows for considering economic aspects; agents may decide to trade with low-trust partners if loss in case of deceit is low. In the paper trust is defined as a subjective probability. This choice

allows trust to be related with risk. In the agent-based simulations of [16], the agent's *trust model* is based on tracing results and the trust update schema proposed in [18]:

$trust_{t+1}(\delta^+)=(1-\delta^+)\ trust_t+\delta^+$ if the experience is positive
$trust_{t+1}(\delta^-)=(1-\delta^-)\ trust_t$ if the experience is negative

where $trust_t$ represents trust after the t transactions. The value of $trust=1$ represents complete trust, $trust=0$ represents complete distrust, and $trust=0.5$ represents complete uncertainty. This model is an asymmetric trust update function, with either completely positive or completely negative experience. A negative experience, or losing something, may have stronger impact than a positive experience, or gaining the same thing. This is known as the endowment effect [1, 26]. δ^+ and δ^- are impact factors of positive and negative experiences respectively. They are related by an endowment coefficient e.

$$\delta^+ = e\,\delta^-,\ 0<e\leq 1$$

This trust update function has the following properties: monotonicity, positive and negative trust extension, and strict positive and negative progression.

For buyers, *trading* entails the trust-or-trace decision. In human interaction, this decision depends on human factors that are not sufficiently well understood to incorporate in a multi-agent system. To not completely disregard these intractable factors, the trust-or-trace decision is modelled as a random process instead of as a deterministic process. The agglomerate of all these intractable factors is called the confidence factor. The distribution involves experience-based trust in the seller, the value ratio of high versus low quality, the cost of tracing, and the buyer's confidence factor.

Tracing reveals the real quality of a commodity. The tracing agent executes the tracing and punishes cheaters as well as traders reselling bad commodities in good faith. The tracing agent only operates on request and requires some tracing fee. Several factors influence the tracing decision to be made after buying a commodity. The most important factors for the current research are the buyer's *trust* in seller and his confidence. Trust is modelled as a subjective evaluation of the probability that the seller would not cheat on the buyer. Confidence reflects the preference of a particular player to trust rather than trace, represented as a value on the interval [0,1]. The other factors are: the satisfaction ratio of the commodity (tracing makes more sense for valuable products than for products with small satisfaction ratio), and the tracing costs which depend on the depth to be traced.

Agent models were validated against the game sessions played by the humans (see [16] for details). Thus, computer simulations were performed for the same game setups with populations of 15 agents: 3 producers, 3 middlemen, 3 retailers, 6 consumers. A set of validation game setups (combinations of free parameters of the agent models) was built to be able to compare output of the computer simulations and the results of the human games. The values of free parameters were selected uniformly from their definition intervals to confirm the models capability to reproduce desired input-output relationships and explore their sensitivities. Some of the highlights of the outcomes are as follows.

Effects of confidence on tracing. The difference in output variables with respect to high and low levels of confidence is not significant for neutral risk-taking agents (i.e.,

agents that evaluate the potential profit equally high as the risk of deception in their buying/selling decision function). For high-risk-taking agents (i.e., agents that prefer potential profit over the risk of deception) the amount of traces decreases for highly confident players and increases for lowly confident players.

Effects of confidence on cheating. The number of cheats is high for highly confident players in games with dominating number of risk-neutral players. Surprisingly, the results show the opposite for risk-taking players: some dishonest sellers do not get traced and punished in highly risk-taking game configurations, so they are encouraged to continue their fraudulent practices.

Effects of confidence on certification and guaranteeing. The number of guarantees is linked with agent's confidence through the concept of the tracing trust. High confidence leads to a lower number of traces, meaning fewer deceptions are discovered and consequently a higher average tracing trust. High tracing trust decreases the seller's risk and thus increases number of guarantees provided by the seller (a guarantee increase seller's costs in case of deception).

In all experiments, effects of risk-taking attitude are consistent: high risk-taking leads to more cheating, less certificates and increased willingness to give guarantees and to rely on them. Differences in risk-taking attitude outweigh changes in other parameters. This result corresponds with observations from human games.

9 Analysis of Trust and Tracing Case Study

The model as introduced in the previous section has been tested in a case study. According to [15], initial trust and honesty of the agents have the strongest influence on the evolution of trust throughout the game. Thus, the following 4 games with homogeneous agents were chosen from the validation set of the game setups to be played using the multi-agent simulation system:

- Distrusting buyers and dishonest sellers
- Distrusting buyers and honest sellers
- Trusting buyers and dishonest sellers
- Trusting buyers and honest sellers

Distrusting buyers have a low initial value of trust (0.1) and develop trust when confronted with honest sellers. Vice versa, trusting buyers have a high initial trust value (1.0) and decrease it through tracing and discovering cheaters in games with dishonest sellers. One additional game was played having only one dishonest agent and in which the rest of the agents are all honest and trusting. Each game played is logged as set of traces generated by each of the agents. Each trace logs the trust and honesty values, and all experiences influencing these. In the logs the trust of an agent is represented by the predicate

has_cheating_trust(AgentA,AgentB,*trust_value*).

This expression means that AgentA has trust value of *trust_value* w.r.t. AgentB. The positive experience influencing the AgentA's trust is recorded using predicate successful_trade(AgentA, AgentB). The predicate has_cheated(AgentA, AgentB) represents the

negative experience (a cheat has been discovered by the tracing agency) that AgentB cheated on AgentA. In a similar way agents log their honesty level that is updated in the same way as their level of trust. The predicate has_honesty(AgentA, *honesty_value*) represents the current value of honesty of AgentA. If the predicate potential_cheat(AgentA) holds, then the value of the honesty of AgentA decreases. If the predicate punishment(AgentA) holds, then the honesty value of AgentA increases.

The game traces were checked against a number of trust properties proposed in Section 5. To this end, the properties of Section 5 have been slightly modified, to make them compatible with the predicates as mentioned above. For example, the properties have been parameterised with a variable for the trusting agent and a variable for the trusted agent, and have been checked for all combinations of agents. The results were as follows:

Positive and negative trust extension. The TTL checker [2] showed that these properties hold for the trust model of the software agents in the Trust and Tracing game. This fact can be also confirmed analytically by the trust update function (see Section 8).

Strict positive and negative monotonic progression. These properties were also confirmed by the TTL checker for all software agents. This is a consequence of the fact that the software agents update their trust using only strictly positive or negative experiences.

Trust flexibility. The automated checks pointed out that this property holds only for a part of the traces for a given value of d. This effect can be explained by the differences in the initial value of trust in the agents across the different games. Furthermore, due to the randomness of the agent's partner selection and the endowment effect (see Section 8) model and the time limits of the games the agents can have insufficient number of trades (experiences) to establish positive trust level .

Positive limit approximation. As explained in Section 7, it makes no sense to check this property against traces of finite length. However, analytically it can be shown that the trust update function converges to the maximum level of trust (1.0) while the agent perceives a continuous sequence of positive experiences.

Limited memory d. According to the TTL checker this property does not hold for the software agents. After each new experience value and performed trust update, the agents still remember all the previous experiences. The history of the experiences is weighted, meaning that the older an experience gets the lower its weight in the current value of trust. In addition, the weight drops exponentially causing the agent to ignore old experiences rapidly.

10 Discussion

This paper focused on the dependence of trust on experiences over time, abstracting from other possible influences, which are assumed as an idealisation to be not present or constant. For trust states as mental states, following literature in Philosophy of Mind such as [20], functional role and representation relation specifications were formalised both in a logical way and in a numerical mathematical way according to a

Dynamical Systems Theory approach (based on integral and differential equations; cf. [25]). It is shown how a backward functional role specification in mathematical terms corresponds to a trust update function in the discrete case and a differential equation in the continuous case and is directly usable in a computational manner for simulation. Moreover, it is shown how a backward representation relation can be obtained (as a repeated application of the functional role specification) by a summation (discrete case) or integration (continuous case) of the (inflating) experiences over time.

Trust may be influenced by experiences of different types. This paper models differences between experiences by mapping them into one overall set of distinct experience 'values'. In addition, more explicit distinctions between different dimensions of experience could be made. Also other cognitive or emotional factors could be integrated, such as the concept of expectation. The work presented in [5, 8, 13, 14] addresses some of these other aspects of trust, which could be integrated. This is left for future work.

In our work we focus on the dynamic properties of the trust models where trust is represented by a single variable. However, we acknowledge the fact that trust is a complex cognitive phenomenon [5] and might be represented by multiple factors or have a more complex relationship with decision making in humans and software agents. To apply our analytical method in such complex cases break down the trust model into single aspect. For each single aspect the proposed dynamic properties can be studied. As an example, this method was applied to the Trust and Tracing case study [14, 15], where different types of trust are used depending on the context of the decision making. The trust-or-trace decision is based on the cheating trust that estimates the subjective probability of truthful behaviour of the opponent while the partner selection decision is based on the negotiation trust, revealing the subjective probability of reaching an agreement with a given partner.

The requirements imposed on models for trust dynamics may depend on individual characteristics of agents; therefore, a variety of models that capture these characteristics may be needed. The approach put forward here enables the explication of these characteristics. Formal specification of both qualitative and quantitative models is supported, based on trust evolution functions and trust update functions. Validation has taken place on the basis of extensive empirical and simulated data in different contexts.

To formalise and analyse dynamic phenomena, often it is implicitly or explicitly claimed that temporal logic, e.g., linear time or branching time temporal logic, is useful; e.g., [3, 9, 10] Other literature claims that the Dynamical Systems Theory, based on differential equations is a suitable approach to dynamics of cognitive phenomena; cf. [25]. In this study it has been found that a number of basic properties for trust dynamics cannot be expressed in standard temporal logic, nor in the form of integral and differential equations. For example, the quite elementary property 'trust monotonicity' that expresses that better experiences lead to more trust is not expressible. The reason for this lack of expressivity is the impossibility to refer to and compare different histories. In the logical language TTL used here, traces are first class citizens; for example, variables and quantifiers can be used over them. In this way explicit reference can be made to histories, and they can be compared.

References

1. Amamor-Boadu, V., Starbird, S.A.: The value of anonimity in supply chain relationships. In: Bremmers, H.J., Omta, S.W.F., Trienekens, J.H., Wubben, E.F.M. (eds.) Dynamics in Chains and Networks, pp. 238–244. Wageningen Academic Publishers, Holland (2004)
2. Bosse, T., Jonker, C.M., Meij, L., van der Sharpanskykh, A., Treur, J.: Specification and Verification of Dynamics in Cognitive Agent Models. In: Proceedings of the Sixth International Conference on Intelligent Agent Technology, IAT'06., pp. 247–254. IEEE Computer Society Press, Los Alamitos (2006)
3. Burgess, J.P.: Temporal logic. In: Gabbay, D.M., Guenther, F. (eds.) Handbook of Philosophical Logic, vol. 2, Reidel, Dordrecht (1984)
4. Castelfranchi, C., Falcone, R.: Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification. In: Demazeau, Y. (ed.) Proceedings of the Third International Conference on Multi-Agent Systems, pp. 72–79. IEEE Computer Society, Los Alamitos (1998)
5. Castelfranchi, C., Falcone, R., Social Trust, A.: Social Trust: A Cognitive Approach. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 55–90. Kluwer Academic Publishers, Dordrecht (2001)
6. Demolombe, R.: To trust information sources: a proposal for a modal logical framework. In: Proceedings of the First International Workshop on Trust, pp. 9–19 (1998)
7. Elofson, G.: Developing Trust with Intelligent Agents: An Exploratory Study. In: Proceedings of the First International Workshop on Trust, pp. 125–139 (1998)
8. Fisher, M.: Temporal Development Methods for Agent-Based Systems. Journal of Autonomous Agents and Multi-Agent Systems 10, 41–66 (2005)
9. Fisher, M., Wooldridge, M.: On the formal specification and verification of multi-agent systems. International Journal of Co-operative Information Systems, IJCIS. In: Huhns, M., Singh, M. (eds.) special issue on Formal Methods in Co-operative Information Systems: Multi-Agent Systems, 6 (1), 37–65 (1997)
10. Galton, A.: Temporal Logic. Stanford Encyclopedia of Philosophy, URL (2003), http://plato.stanford.edu/entries/logic-temporal/#2
11. Galton, A.: Operators vs Arguments: The Ins and Outs of Reification. Synthese 150, 415–441 (2006)
12. Gambetta, D.: Trust. Basil Blackwell, Oxford (1990)
13. Grandison, T., Sloman, M.: A Survey of Trust in Internet Applications. IEEE Communications Surveys (2000)
14. Jonker, C.M., Meijer, S., Tykhonov, D., Verwaart, T.: Agent-based Simulation of the Trust and Tracing Game for Supply Chains and Networks. Technical Report, Delft University of Technology
15. Jonker, C.M., Meijer, S., Tykhonov, D., Verwaart, D.: Modelling and Simulation of Selling and Deceit for the Trust and Tracing Game. In: Castelfranchi, C., Barber, S., Sabater, S., Singh, M. (eds.) Proceedings of the Trust in Agent Societies Workshop, pp. 78–90. Springer, Heidelberg (2005)
16. Jonker, C.M., Schalken, J.J.P., Theeuwes, J., Treur, J.: Human Experiments in Trust Dynamics. In: Jensen, C., Poslad, S., Dimitrakos, T. (eds.) iTrust 2004. LNCS, vol. 2995, pp. 206–220. Springer, Heidelberg (2004)
17. Jonker, C.M., Treur, J.: Formal analysis of models for the dynamics of trust based on experiences. In: Garijo, F.J., Boman, M. (eds.) MAAMAW 1999. LNCS, vol. 1647, pp. 221–232. Springer, Heidelberg (1999)

18. Jøsang, A., Presti, S.: Analysing the Relationship between Risk and Trust. In: Jensen, C., Poslad, S., Dimitrakos, T. (eds.) Trust 2004. LNCS, vol. 2995, pp. 135–145. Springer, Heidelberg (2004)
19. Kim, J.: Philosophy of Mind. Westview Press, Boulder (1996)
20. Lewis, D., Weigert, A.: Social Atomism, Holism, and Trust. In: Sociological Quarterly, pp. 455–471 (1985)
21. Marsh, S.: Trust and Reliance in Multi-Agent Systems: a Preliminary Report. In: Proc. of the Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'92, Rome (1992)
22. Marsh, S.: Trust in Distributed Artificial Intelligence. In: Castelfranchi, C., Werner, E. (eds.) MAAMAW 1992. LNCS, vol. 830, pp. 94–112. Springer, Heidelberg (1994)
23. Meijer, S., Hofstede, G.J.: The Trust and Tracing game. In: Proceedings of 7th Int. workshop on experiential learning. IFIP WG 5.7 SIG conference, Aalborg, Denmark (2003)
24. Port, R.F., van Gelder, T. (eds.): Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge, Mass (1995)
25. Rabin, M.: Psychology and Economics. Journal of Economic Literature 36, 11–46 (1998)
26. Ramchurn, S.D., Hunyh, D., Jennings, N.R.: Trust in Multi-Agent Systems, Knowledge Engineering Review (2004)