# Cognitive and Neural Modeling of Dynamics of Trust in Competitive Trustees

Mark Hoogendoorn, S. Waqar Jaffry, and Jan Treur
*Vrije Universiteit Amsterdam, Department of Artificial Intelligence De Boelelaan 1081a,*
*1081 HV Amsterdam, The Netherlands.*
{mhoogen, swjaffry, treur}@few.vu.nl

**Abstract.** Trust dynamics can be modeled in relation to experiences. In this paper two models to represent human trust dynamics are introduced, namely a model on a cognitive level and a neural model. These models each include a number of parameters, providing the possibility to express certain relations between trustees. The behavior of each of the models is further analyzed by means of simulation experiments and formal verification techniques. Thereafter, both models have been compared to see whether they can produce patterns that are comparable. As each of the models has its own specific set of parameters, with values that depend on the type of person modeled, such a comparison is nontrivial. To address this, a special comparison approach is introduced, based on mutual mirroring of the models in each other. More specifically, for a given parameter values set for one model, by an automated parameter estimation procedure the most optimal values for the parameter values of the other model are determined in order to show the same behavior. Roughly spoken the results are that the models can mirror each other up to an accuracy of around 90%.

**Keywords:** trust dynamics, cognitive, neural, comparison, parameter tuning.

## 1. Introduction

Nowadays, more and more ambient systems are being deployed to support humans in an effective way (Aarts, Harwig, and Schuurmans, 2001; Aarts, Collier, van Loenen, and Ruyter, 2003; Riva, Vatalaro, Davide, and Alcaniz, 2005). An example of such an ambient system is a personal agent that monitors the behaviour of a human executing certain complex tasks, and gives dedicated support for this. Such support may include advising the use of a particular information source, system or agent to enable proper execution of the task, or even involving such a system or agent pro-actively. In order for these personal agents to be accepted and useful, the personal agent should be well aware of the habits and preferences of the human. If a human for example for good reasons dislikes using a particular system or agent, and there are several alternatives available that are more preferred, the personal agent would not be supporting effectively if it would advise, or even pro-actively initiate, the disliked option.

An aspect that plays a crucial role in giving such tailored advice is to represent the trust levels the human has for certain options. Knowing these trust values allows the personal assistant to reason about these levels, and give the best possible support that is in accordance with the habits and preferences of the human. Since there would be no problem in case there is only one way of supporting the human, the problem of selecting the right support method only occurs in case of substitutable options. Therefore, a notion of relative trust in these options seems more realistic than having a separate independent trust value for each of these options. For instance, if three systems or agents can contribute X, and two of them perform bad, whereas the third performs pretty bad as well, but somewhat better in than the others, trust in that third option may still be a relatively high since in the context of the other options it is the best alternative. The existing trust models do however not explicitly handle such relative trust notions (see e.g. Falcone and Castelfranchi, 2004; Jonker and Treur, 1999; Marx and Treur).

In this paper, a cognitive and a neural model are presented that address the dynamics of trust, including the aforementioned notion of relative trust and particular other personality characteristics. Both models are evaluated using simulation experiments and formal verification techniques.

The first model, representing trust on a cognitive level, takes into account two main functional properties of trust states, which define the *causal or functional role* of a trust state as cognitive state, as put forward in (Jonker and Treur, 2003):

(1) A trust state results from accumulation of experiences over time

(2) Trust states affect decision making by choosing more trusted options above less trusted options

The second model of trust dynamics is based on neurological principles. In this model, theories on the interaction between affective and cognitive states (see e.g., Eich, Kihlstrom, Bower, Forgas, and Niedenthal, , 2000; Forgas, Laham, and Vargas, 2005; Forgas, Goldenberg, and Unkelbach, 2009; Niedenthal, 2007; Schooler and Eich, 2000; Winkielman, Niedenthal, and Oberman, 2009) are modeled on a neurological level as well by using theories on the embodiment of emotions as described, for example, in (Winkielman, Niedenthal, and Oberman, 2009; Damasio, 1994; Damasio, 1996; Damasio, 1999; Damasio, 2003), it is described how trust dynamics relates to experiences with (external) sources, both from a cognitive and affective perspective. More specifically, in accordance with, for example (Damasio, 1999; Damasio, 2003), for feeling the emotion associated to a mental state, a converging *recursive body loop* is assumed. In addition, based on *Hebbian learning* (cf. Hebb, 1949; Bi and Poo, 2001; Gerstner, and Kistler, 2002) for the strength of the connections to the emotional responses, an adaptation process is introduced, inspired by the Somatic Marker Hypothesis (Damasio, 1994; Damasio, 1996).

Being described on a different level, each of the models includes specific set of parameters for cognitive and neurological characteristics of the person being modeled. As the set of parameters of these models have no known connection with each other, and the behavior of such models strongly depends on the values for such parameters, a direct comparison between the models is impossible. Therefore a comparison between the models is made in a more indirect way, by mutual *mirroring* them in each other. This mirroring approach uses any set of values that is assigned to the parameters for one of the models to obtain a number of simulation traces. These simulation traces are approximated by the second model, based on automated parameter estimation. The error for this approximation is considered as a comparison measure. The mirroring is applied in two directions, and also back and forth sequentially by using the estimated parameter values for the second model to estimate new parameter values for the first.

In the paper, first in Section 2 the cognitive model for trust dynamic is described, and in Section 3 simulation results of this model. Section 4 presents a formal analysis of the model. In Section 5 the neural model is presented, with simulation results discussed in Section 6. Section 7 presents a formal analysis of the neural model. In Section 8 the mirroring approach for comparison of models and the automated parameter estimation method are discussed. Finally, Section 9 is a discussion.

## 2 A Cognitive Model for Relative Trust

This section proposes a cognitive model that caters the dynamics of a human's trust on competitive trustees. In this model trust of the human on a trustee depends on the relative experiences with the trustee in comparison to the experiences from all of the competitive trustees. The model defines the total trust of the human as the difference between positive trust and negative trust (distrust) on the trustee. It includes personal human characteristics like trust decay, flexibility, and degree of autonomy (context-independence) of the trust. Figure 1 shows the dynamic relationships in the proposed model.

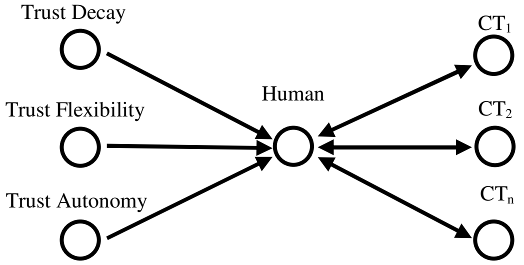


**Fig. 1.** Trust-based interaction with *n* competitive trustees

In this model it is assumed that the human is bound to request one of the available competitive trustees at each time step. The probability of the human's decision to request one of the trustees $\{CT_1, CT_2, . . . CT_n\}$ at time $t$ is based on the trust value $\{T_1, T_2, . . . T_n\}$ for each $CT_i$ respectively at time $t$. In the response of the human's request $CT_i$ gives experience value ($E_i(t)$) from the set $\{-1, 1\}$ which means a negative and positive experience respectively. This experience is used to update the trust value for the next time point. Besides $\{-1, 1\}$ the experience value can also be 0, indicating that $CT_i$ gives no experience to the human at time point $t$. First, the parameters that characterize human behavior in this model are explained. Thereafter, the model itself which incorporates these parameters is expressed. Simulation results are shown using the model with various parameter settings. Finally, a formal mathematical analysis of the model is presented.

## 2.1 Parameters Characterizing Individual Differences between Humans

To tune the model to specific personal human characteristics a number of parameters are used.

**Flexibility $\beta$** The personality attribute called trust flexibility ($\beta$) is a number between [*0, 1*] that represents in how far the trust level at time point *t* will be adapted when human has a (positive or negative) experience with a trustee. If this factor is high then the human will give more weight to the experience at $t+\Delta t$ than the already available trust at *t* to determine the new trust level for $t+\Delta t$ and vice versa.

**Trust Decay $\gamma$** The human personality attribute called trust decay ($\gamma$) is a number between [*0, 1*] that represents the rate of trust decay of the human on the trustee when there is no experience. If this factor is high then the human will forget soon about past experiences with the trustee and vice versa.

**Autonomy $\eta$** The human personality attribute called autonomy ($\eta$) is a number between [*0, 1*] that indicates in how far trust is determined independent of trust in other options. If the number is high, trust is (almost) independent of other options.

**Initial Trust** The human personality attribute called initial trust indicates the level of trust assigned initially to a trustee.

## 2.2 A Cognitive Model for Relative Trust

The model is composed from two models: one for the positive trust, accumulating positive experiences, and one for negative trust, accumulating negative experiences. The approach of taking positive and negative trust separately at the same time to measure total trust is similar to the approaches taken in literature for degree of belief and disbelief (Shortliffe and Buchanan, 1975) and (Luger and Stubblefield, 1998). Both negative and positive trusts are a number between [*0, 1*]. While human total trust at $CT_i$ on any time point t is the difference of positive and negative trust at $CT_i$ at time *t*.

Here first the positive trust is addressed. The human's relative positive trust of $CT_i$ at time point *t* is based on a combination of two parts: the *autonomous* part, and the *context-dependent* part. For the latter part an important indicator is the human's relative positive trust of $CT_i$ at time point *t* (denoted by $\tau_i^+(t)$): the ratio of the human's trust of $CT_i$ to the average human's trust on all options at time point *t*. Similarly an indicator for the human's relative negative trust of $CT_\text{i}$ at time point *t* (denoted by $\tau_i^-(t)$) is the ratio between human's negative trust of the option $CT_i$ and the average human's negative trust on all options at time point *t*. These are calculated as follows:

$$\tau_i^+(t) = \frac{T_i^+(t)}{\sum_{j=1}^n T_j^+(t)/n} \text{ and } \tau_i^-(t) = \frac{T_i^-(t)}{\sum_{j=1}^n T_j^-(t)/n}$$

Here the denominators $\sum_{j=1}^n T_j^+(t)/n$ and $\sum_{j=1}^n T_j^-(t)/n$ express the average positive and negative trust over all options at time point *t* respectively. The context-dependent part was designed in such a way that when the positive trust is above the average, then upon each positive experience it gets an extra increase, and when it is below average it gets a decrease. This models a form of competition between the different options. The principle used is a variant of a 'winner takes it all' principle, which for example is sometimes modeled by mutually inhibiting neurons representing the different options. This principle has been modeled by basing the change of trust upon a positive experience on $\tau_i^+(t) - 1$, which is positive when the positive trust is above average and negative when it is below average. To normalize, this is multiplied by a factor $T_i^+(t)*(1 - T_i^+(t))$. For the autonomous part the change upon a positive experience is modeled by $1 - T_i^+(t)$. In this formulation $\eta$ indicates in how far the human is autonomous or context-dependent in trust attribution therefore a weighted sum is taken with weights $\eta$ and $1\text{-}\eta$ respectively. Therefore, using the parameters defined in above $T_i^+(t+\Delta t)$ is calculated using the following equations. Note that here the competition mechanism is incorporated in a dynamical systems approach where the values of $\tau_i^+(t)$ have impact on the change of positive trust over time. Followings are the equations when $E_i(t)$ is *1*, *0* and *-1* respectively.

$$\begin{aligned} T_i^+(t+\Delta t) = {} & T_i^+(t) + \\ & \beta \begin{bmatrix} \eta * (1 - T_i^+(t)) + \\ (1-\eta) * (\tau_i^+(t) - 1) * T_i^+(t) * \ (1 - T_i^+(t)) \end{bmatrix} \\ & * \Delta t \\ & \text{when } E_i(t) = 1 \\ T_i^+(t+\Delta t) = {} & T_i^+(t) - \gamma * T_i^+(t) * \Delta t \\ & \text{when } E_i(t) = 0 \\ T_i^+(t+\Delta t) = {} & T_i^+(t) \\ & \text{when } E_i(t) = -1 \end{aligned}$$

Notice that here in the case of negative experience positive trust is kept constant to avoid doubling the effect over all trust calculation as negative experience is accommodated fully in the negative trust calculation. In one formula this is expressed by:

$$\begin{aligned} T_i^+(t+\Delta t) = T_i^+(t) + [\beta * [\eta * \left(1 - T_i^+(t)\right) \\ + (1-\eta) * (\tau_i^+(t) - 1) * T_i^+(t) \\ * \left(1 - T_i^+(t)\right)] * E_i(t) \\ * (E_i(t) + 1)/\,2 - \gamma * T_i^+(t) \\ * (E_i(t) + 1) * (E_i(t) - 1)] \ * \Delta t \end{aligned}$$

In differential equation form this can be reformulated as:

$$\begin{aligned} \frac{dT_i^+(t)}{dt} = \beta * \ [\eta * \left(1 - T_i^+(t)\right) + (1-\eta) \\ * (\tau_i^+(t) - 1\,) * T_i^+(t) \\ * \left(1 - T_i^+(t)\right)] * E_i(t) \\ * (E_i(t) + 1)/\,2 - \gamma * T_i^+(t) \\ * (E_i(t) + 1) * (E_i(t) - 1) \end{aligned}$$

Notice that this is a system of *n* coupled differential equations; the coupling is realized by $\tau_i^+(t)$ which includes the sum of the different trust values for all j. Similarly, for negative trust followings are the equations when $E_i(t)$ is *-1*, *0* and *1* respectively.

$$\begin{aligned} T_i^-(t+\Delta t) = \quad & T_i^-(t) + \ \beta \\ & * \left[\eta * \left(1 - T_i^-(t)\right) + \ (1-\eta)\right. \\ & * (\tau_i^-(t) - 1) * T_i^-(t) \\ & * \ \left(1 - T_i^-(t)\right)] * \Delta t \\ & \text{when } E_i(t) = -1 \\ T_i^-(t+\Delta t) = \ & T_i^-(t) - \gamma * T_i^-(t) * \Delta t \\ & \text{when } E_i(t) = \ 0 \\ T_i^-(t+\Delta t) = \ & T_i^-(t) \\ & \text{when } E_i(t) = \ 1 \end{aligned}$$

In one formula this is expressed as:

$$\begin{aligned} T_i^-(t+\Delta t) = T_i^-(t) + [\beta * [\eta * \left(1 - T_i^-(t)\right) \\ + (1-\eta) * (\tau_i^-(t) - 1) * T_i^-(t) \\ * \left(1 - T_i^-(t)\right)] * E_i(t) \\ * (E_i(t) - 1)/\,2 - \gamma * T_i^-(t) \\ * (E_i(t) + 1) * (E_i(t) - 1)] \ * \Delta t \end{aligned}$$

In differential equation form this can be reformulated as:

$$\begin{aligned} \frac{dT_i^-(t)}{dt} = \beta * \ [\eta * \left(1 - T_i^-(t)\right) + (1-\eta) \\ * (\tau_i^-(t) - 1\,) * T_i^-(t) \\ * \left(1 - T_i^-(t)\right)] * E_i(t) \\ * (E_i(t) - 1)/\,2 - \gamma * T_i^-(t) \\ * (E_i(t) + 1) * (E_i(t) - 1) \end{aligned}$$

Notice that this again is a system of *n* coupled differential equations but not coupled to the system for the positive case described above.

**Combining positive and negative trust.** The notions of positive and negative relative trust can be combined into a single overall relative trust. Hereby, the human's total trust $T_i(t)$ of $CT_i$ at time point *t* is a number between *[-1, 1]* where *-1* and *1* represent minimum and maximum values of the trust respectively. It is the difference of the human's positive and negative trust of $CT_i$ at time point *t*:

$$T_i(t) = T_i^+(t) - T_i^-(t)$$

In particular, also the human's initial total trust of $CT_i$ at time point *0* is $T_i(0)$ which is the difference of human's initial trust $T_i^+(0)$ and distrust $T_i^-(0)$ in $CT_i$ at time point *0.*

**Decision making model.** The final step is to model the decision making of the human. As the human's total trust is a number in the interval *[-1, 1]*, to calculate the *request probability* to request $CT_i$ at time point *t* ($RP_i(t)$) the human's total trust $T_i(t)$ is first projected at the interval *[0, 2]* and then normalized as follows;

$$RP_i(t) = \frac{T_i(t) + 1}{\sum_{j=1}^{n}\left(T_j(t) + 1\right)}$$

## 3 Simulations for the Cognitive Model

In this section a number of simulation results are presented to describe behavior of the model designed in section 2. Here human's total trust on the three competitive Information Agents (IA's) is calculated. It is assumed that the human is bound to request one of the available competitive information agents at each time step. The probability of the human's decision to request one of the information agents {$IA_1$, $IA_2$, $IA_3$} at time *t* is based on the human's total trust with each information agent respectively at time *t* {$T_1(t)$, $T_2(t)$, $T_3(t)$} (i.e. the equation shown in section 2.4). In response of the human's request for information the agent gives an experience value $E_i(t)$.

### 3.1 Relativeness

The first experiment described is conducted to observe different aspects, including the relativeness attribute of the model (see Figure 2). In the Figure, the x-axis represents time, whereas the y-axis represents the trust value for the various information agents. The configurations taken into the account are as shown in Table 1.

**Table 1.** Parameter values to analyze the dynamics of relative trust with the change in IAs responses.

| Attribute | Symbol | Value |
|---|---|---|
| Trust Decay | $\gamma$ | 0.01 |
| Autonomy | $\eta$ | 0.25 |
| Flexibility | $\beta$ | 0.75 |
| Time Step | $\Delta t$ | 0.10 |
| Initial Trust and Distrust of $\{IA_1,IA_2,IA_3\}$ | $T_1^+(0)$, $T_2^+(0)$, $T_3^+(0)$,<br>$T_1^-(0)$, $T_2^-(0)$, $T_3^-(0)$ | 0.50, 0.50, 0.50,<br>0.50, 0.50, 0.50 |

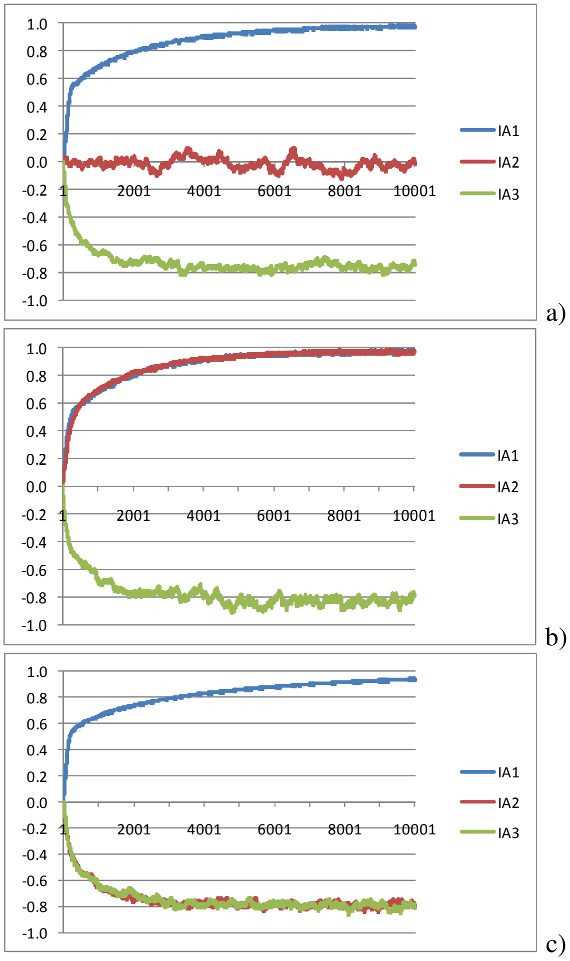


**Fig. 2.** Model Dependence on amount of positive response from IAs: a) Information Agents $IA_1$, $IA_2$, $IA_3$ give experience positive, random (equal probability to give a positive or negative experience), negative respectively on each request by the Human respectively. b) Information Agents $IA_1$, $IA_2$, $IA_3$ give experience positive, positive, negative on each request by the Human respectively. c) Information Agents $IA_1$, $IA_2$, $IA_3$ give experience positive, negative, negative on each request by the Human respectively.

It is evident from above graphs that the information agent who gives more positive experience gets more relative trust than the others, which can be considered a basic property of trust dynamics (trust monotonicity) (Marx and Treur, 2001) and (Jonker and Treur, 1999).

### 3.2 Trust Decay

This second experiment, shown in Figure 3, is configured to observe the change in the total trust in relation to change in the trust decay attribute $\gamma$ of the human. The configurations taken into the account are as shown in Table 2.

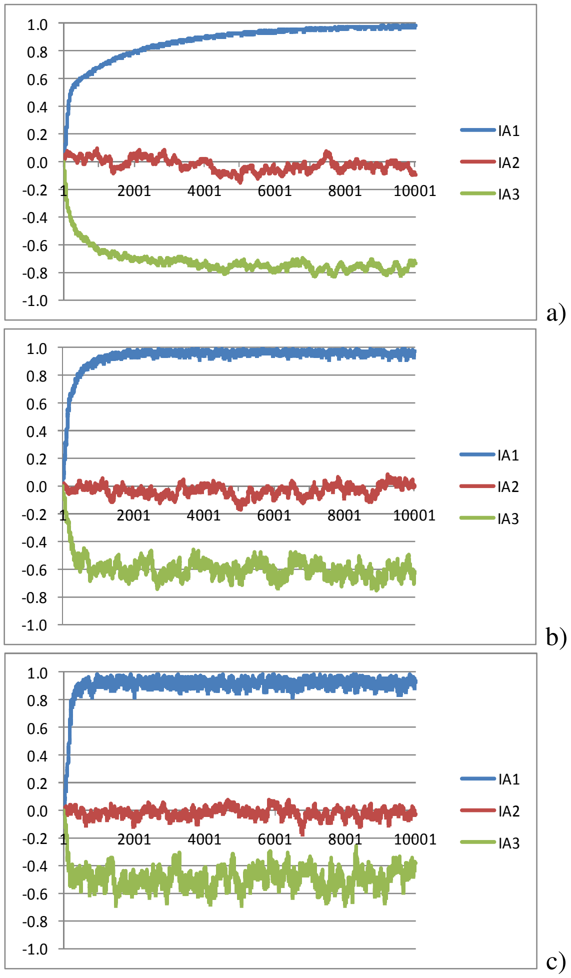


**Fig. 3.** Model Dependence on Trust Decay: a) $\gamma$=0.01. b) $\gamma$=0.05. c) $\gamma$=0.10.

**Table 2.** Parameter values to analyze the dynamics of relative trust with the change in trust decay ($\gamma$).

| Attribute | Symbol | Value |
|---|---|---|
| Experience {$IA_1$, $IA_2$, $IA_3$} | $E_1$, $E_2$, $E_3$ | 1, random, -1 |
| Autonomy | $\eta$ | 0.25 |
| Flexibility | $\beta$ | 0.75 |
| Time Step | $\Delta t$ | 0.10 |
| Initial Trust and Distrust of {$IA_1$,$IA_2$,$IA_3$} | $T_1^+(0)$, $T_2^+(0)$, $T_3^+(0)$,<br>$T_1^-(0)$, $T_2^-(0)$, $T_3^-(0)$ | 0.50,0.50,0.50,<br>0.50,0.50,0.50 |

In these cases also the information agent who gives more positive experience gets more relative trust than the others. Furthermore, if the trust decay is higher, then the trust value drops rapidly on no experience (see Figure 3c; more unsmooth fringes of the curve).

### 3.3 Flexibility of Trust

This experiment is configured to observe the change in the total trust with the change in the human's flexibility of the trust (see Figure 4). Configurations taken into the account are shown in Table 3.

**Table 3.** Parameter values to analyze the dynamics of relative trust with the change in flexibility ($\beta$).

| Attribute | Symbol | Value |
|---|---|---|
| Experience {$IA_1$, $IA_2$, $IA_3$} | $E_1$, $E_2$, $E_3$ | 1, random, -1 |
| Trust Decay | $\gamma$ | 0.01 |
| Autonomy | $\eta$ | 0.25 |
| Time Step | $\Delta t$ | 0.10 |
| Initial Trust and Distrust of {$IA_1$,$IA_2$,$IA_3$} | $T_1^+(0)$, $T_2^+(0)$, $T_3^+(0)$,<br>$T_1^-(0)$, $T_2^-(0)$, $T_3^-(0)$ | 0.50, 0.50, 0.50,<br>0.50, 0.50, 0.50 |

In these cases again the information agent who gives more positive experience gets more human's relative trust then the others. Furthermore as the values of the $\beta$ decrease the rate of change of the trust also decrease. In Figure 4c, $\beta$=0 which means that trust does not change on experiences at all, so the initial values retain for experiences from the information agents hence trust value remains stable. Finally in the Figure 4d as initial values of the total trust are taken $T_1(0)=1$, $T_2(0)=0$ and $T_3(0)=-1$ instead of $T_1(0)=0$, $T_2(0)=0$ and $T_3(0)=0$, so the total trust decays due to the trust decay factor and becomes stable after a specific time span.

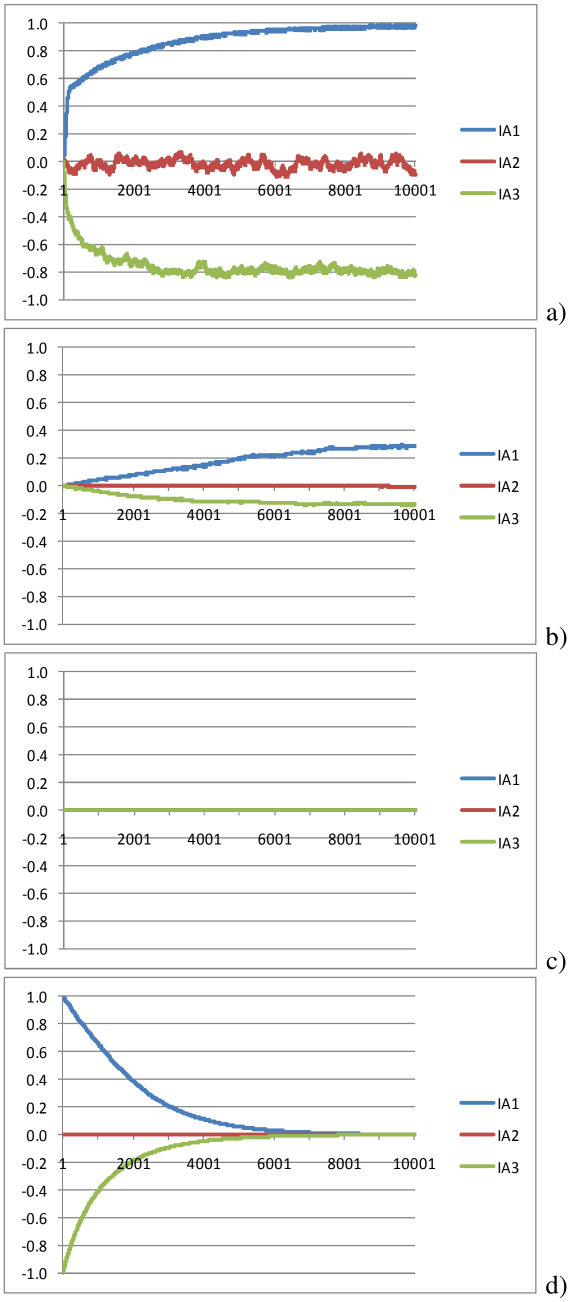


**Fig. 4.** Model Dependence on Trust Flexibility: a) $\beta$=1, b) $\beta$=0.01, c) $\beta$=0.00, d) $\beta$=0.00 and $T_1(0)=1$, $T_2(0)=0$, $T_3(0)=-1$.

### 3.4 Autonomy of Trust

This experiment (see Figure 5) is configured to observe the change in the human trust with the change in the human's autonomy for the total trust calculation. Configurations taken into the account are shown in Table 4.

**Table 4.** Parameter values to analyze the dynamics of relative trust with the change in autonomy (*η*).

| Attribute | Symbol | Value |
|---|---|---|
| Experience $\{IA_1, IA_2, IA_3\}$ | $E_1, E_2, E_3$ | 1, random, -1 |
| Trust Decay | γ | 0.01 |
| Flexibility | β | 0.75 |
| Time Step | Δt | 0.10 |
| Initial Trust and Distrust of $\{IA_1,IA_2,IA_3\}$ | $T_1^+(0), T_2^+(0), T_3^+(0),$<br>$T_1^-(0), T_2^-(0), T_3^-(0)$ | 0.50, 0.50, 0.50,<br>0.50, 0.50, 0.50 |

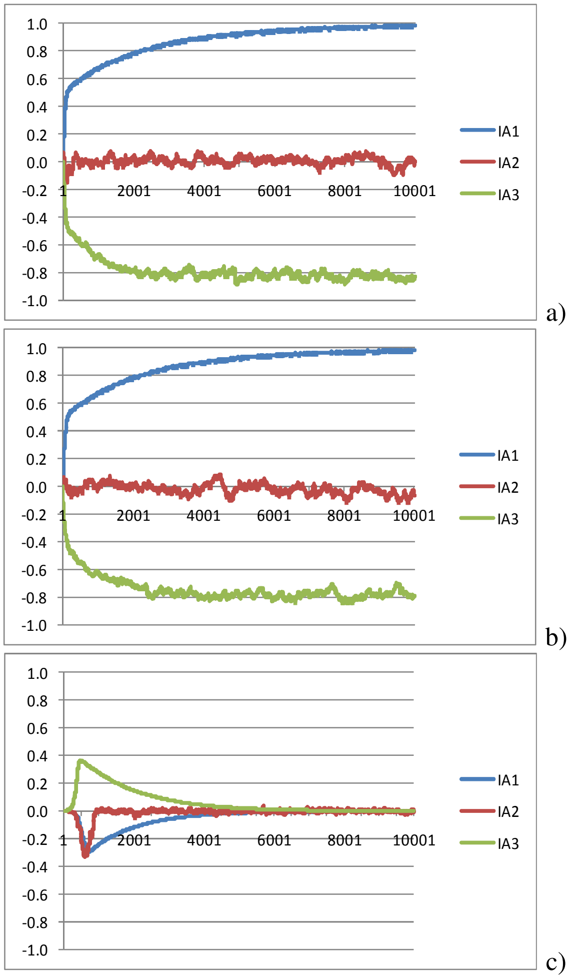


**Fig. 5.** Model Dependence on Trust Autonomy: a) *η*=1.0, b) *η*=0.50, c) *η*=0.00.

In these cases also the information agent who gives more positive experience gets more relative trust then the others. Furthermore as the values of the η decrease the human weights the relative part of the trust more than the autonomous trust. In Figure 5c, *η=0* which means that the human does not take into account the autonomous trust. This gives unstable patterns that are extremely sensitive to the initial conditions of the system. The example graph shown is just one of these patterns.

### 3.5 Initial Trust and Distrust

This experiment is configured to observe the change in the total trust with the change in the human's initial trust and distrust ($T^+_i(0)$, $T^-_i(0)$) on information agents (see Figure 6). Configurations taken into the account are shown in Table 5.

**Table 5.** Parameter values to analyze the dynamics of relative trust with the change in initial trust.

| Attribute | Symbol | Value |
|---|---|---|
| Experience $\{IA_1, IA_2, IA_3\}$ | $E_1, E_2, E_3$ | 1, random, -1 |
| Trust Decay | γ | 0.01 |
| Autonomy | η | 0.25 |
| Flexibility | β | 0.75 |
| Time Step | Δt | 0.10 |

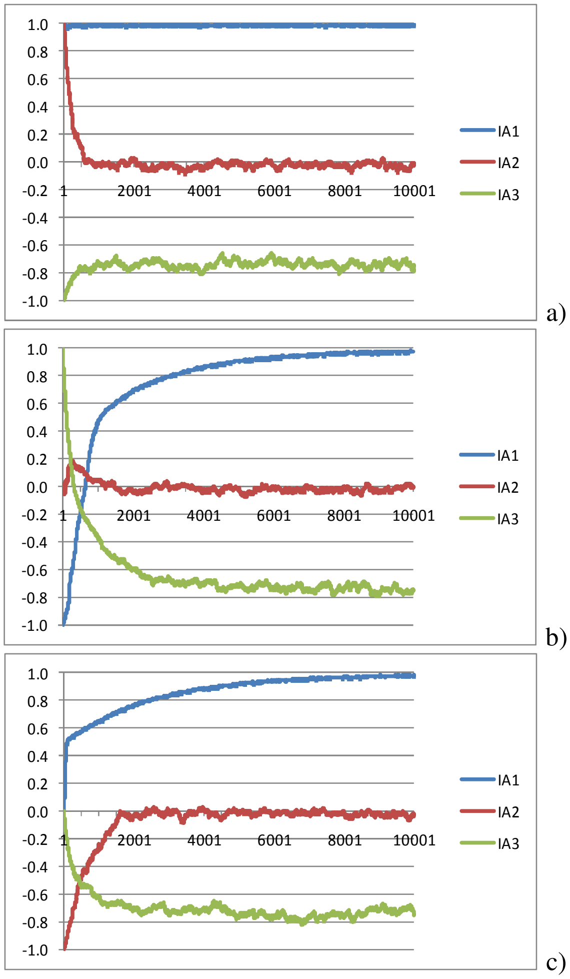


**Fig. 6.** Model Dependence on Initial Trust $\{T_1(0), T_2(0), T_3(0)\}$: a) 1, 1, -1. b) -1, 0, 1. c) 0, -1, 0.

It is observed from the above graphs that the final outcome of the trust is not very sensitive for the initial values.

### 3.6 Dynamics of Trust in Different Cultures

The degree of reliability of available information sources may strongly differ in different types of societies or cultures. In some types of societies it may be exceptional when an information source provides 10% or more false information, whereas in other types of societies it is more or less normal that around 50% of the outcomes of information sources is false. If the positive experiences percentage given by the information agents varies significantly, then the total relative trust of the human on these information agents may differ as well.

This case study was designed to study dynamics of the human's trust on information agents in different cultures with respect to the percentages of the positive experiences they provide to the human. A main question is whether in a culture where most information sources are not very reliable, the trust in a given information source is higher than in a culture where the competitive information sources are more reliable. Here cognitive model of relative trust described in section 2 is used for simulation purposes. Cultures are named with respect to percentage of the positive experiences provided by the information agents to the human as shown in Table 6 and other experimental configurations in Table 7.

**Table 6.** Classification of Human Cultures with respect to the Positive Experiences given by the IAs.

| Culture Name | Percentage of the positive experiences by the information agents $\{IA_1, IA_2, IA_3\}$ |
|---|---|
| A | 100, 99, 95 |
| B | 50, 40, 30 |
| C | 10, 0, 0 |
| D | 0,0,0 |

**Table 7.** Parameter values to analyze the Relative Trust Dynamics in different Cultures.

| Attribute | Symbol | Value |
|---|---|---|
| Trust Decay | $\gamma$ | 0.01 |
| Autonomy | $\eta$ | 0.25 |
| Flexibility | $\beta$ | 0.75 |
| Time Step | $\Delta t$ | 0.10 |
| Initial Trust and Distrust of $\{IA_1, IA_2, IA_3\}$ | $T_1^+(0)$, $T_2^+(0)$, $T_3^+(0)$, $T_1^-(0)$, $T_2^-(0)$, $T_3^-(0)$ | 0.50,0.50,0.50<br>0.50,0.50,0.50 |

Simulation results for the dynamics of the relative trust for the cultures mentioned in Table 6 are shown in Figure 7.

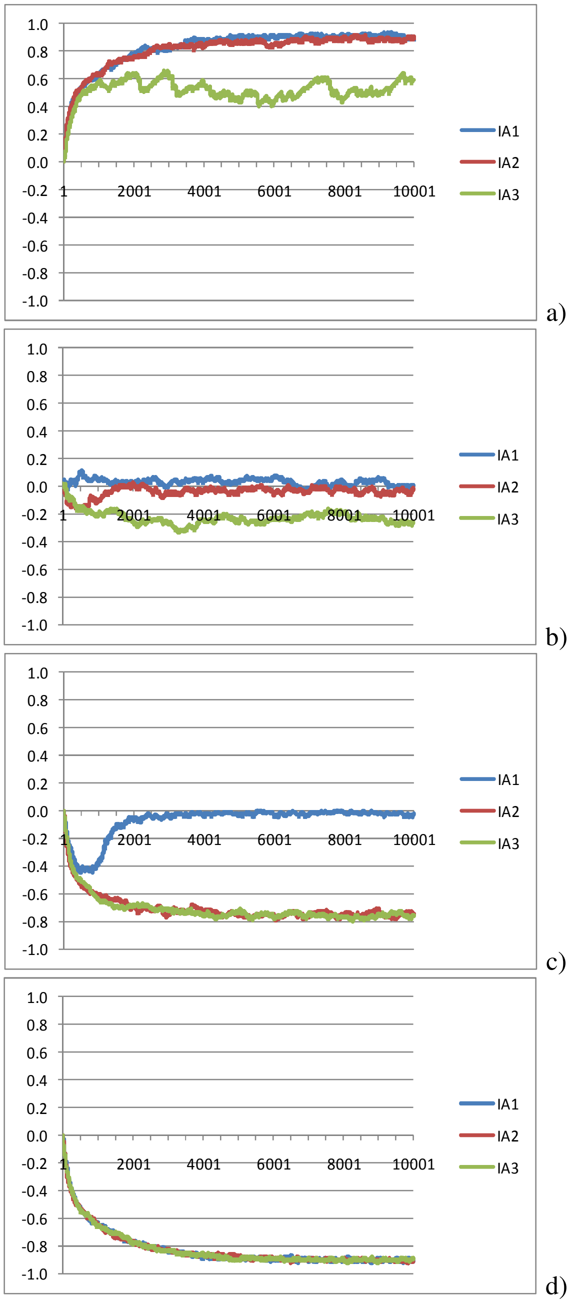


**Fig. 7.** Dynamics of Relative Trust in Different Cultures. a) Culture A, b) Culture B, c) Culture C, d) Culture D

From Figure 7 it can be concluded that in every culture whatever relative percentage of the positive experiences may be (except when all information agent give negative experiences all of the time (see Figure 7d), the information agent that gives more positive experiences to the human gains more trust. Furthermore, the information agent that gives more positive experiences at least secure neutral trust (*T(t)=0*) in the long run, even the percentage of positive experiences is very low (see Figure 7c).

## 4 Analysis of the Cognitive Model

In this section a mathematical analysis is presented of the change in trust upon positive and negative experiences. In Section 2 the differential equation form of the model for positive trust was formulated as:

$$\begin{aligned}\frac{dT_i^+(t)}{dt} = \beta * \ [\eta * \left(1 - T_i^+(t)\right) - (1 - \eta) \\ * (1 - \ \tau_i^+(t)\ ) * T_i^+(t) \\ * \left(1 - T_i^+(t)\right)] \ * E_i(t) \\ * \left(1 + E_i(t)\right) / \ 2 - \gamma * T_i^+(t) \\ * \left(1 + E_i(t)\right) * \left(1 - E_i(t)\right)\end{aligned}$$

with

$$\tau_i^+(t) = \frac{T_i^+\ (t)}{\sum_{j=1}^{n} T_j^+(t) \Big/ n}$$

One question that can be addressed is when for a given time point *t* an equilibrium occurs, i.e. under which conditions trust does not change at time point *t*. Another question is under which circumstances trust will increase at *t*, and under which it will decrease. As the experience function $E_i(t)$ is given by an external scenario, these questions have to be answered for a given value of this function. So, three cases are considered:

**Case 1:** $E_i(t) = 1$
In this case the differential equation can be simplified to

$$\begin{aligned}\frac{dT_i^+(t)}{dt} = \beta * [\eta * \left(1 - T_i^+(t)\right) - \ (1 - \eta) \\ * (1 - \ \tau_i^+(t)\ ) * T_i^+(t) \\ * \left(1 - T_i^+(t)\right)]\end{aligned}$$

$$\frac{dT_i^+(t)}{dt} = \beta * \ [\eta \ - \ (1 - \eta) * (1 - \frac{T_i^+\ (t)}{\sum_{j=1}^{n} T_j^+(t) \Big/ n})$$

$$* T_i^+(t)] * \left(1 - T_i^+(t)\right)$$

It follows that $\frac{dT_i^+(t)}{dt} \geq 0$ if and only if

$$[\eta \ - \ (1 - \eta) * (1 - \frac{T_i^+\ (t)}{\sum_{j=1}^{n} T_j^+(t) \Big/ n}) * T_i^+(t)\ ] \qquad \geq 0$$

or

$T_i^+(t) = 1.$

For $T_i^+(t) < 1$ this is equivalent to (with *S(t)* $= \sum_{j=1}^{n} T_j^+(t)$):

$$(1 - \eta) * \left(1 - \frac{T_i^+\ (t)}{S(t)/n}\right) * T_i^+(t)\ ] \qquad \leq \eta$$

$$(1 - \eta) * (S(t) - n\, T_i^+(t)\ ) * T_i^+(t)\ ] \qquad \leq \eta\, S(t)$$

$$S(t) T_i^+(t)\ - n\, T_i^+\ (t)^2 \qquad \leq \eta$$

$$S(t)/(1 - \eta)$$

$$n\, T_i^+(t)^2 - S(t) T_i^+(t) + \eta\, S(t)/(1 - \eta) \qquad \geq 0$$

This quadratic expression in $T_i^+(t)$ has no zeros when the discriminant $S(t)^2 - \frac{4nS(t)\eta}{1-\eta}$ is negative:

$$S(t)^2 - \frac{4nS(t)\eta}{1-\eta} < 0 \ \Leftrightarrow\ S(t) * (\, S(t) - \frac{4n\eta}{1-\eta}) < 0$$
$$\Leftrightarrow\ \ 0 < \ S(t)/n < \frac{4\eta}{1-\eta}$$

When $\eta > 0.2$ then $1/\eta < 5$ and therefore $1/\eta - 1 < 4$, hence $(1\text{-}\eta\ ) / \eta < 4$ which can be reformulated as $\frac{4\eta}{1-\eta} > 1$. As $S(t)/n \leq 1$, this shows that for $\eta > 0.2$ as long as *S(t)* is positive, the discriminant is always negative, and therefore upon a positive experience there will always be an increase. When *S(t) = 0*, which means all trust values are *0*, no change occurs. For the case the discriminant is ≥0, i.e., $S(t)/n \geq \frac{4\eta}{1-\eta}$ then the quadratic equation for $T_i^+(t)$ has two zeros symmetric in *S(t):*

$$T_i^+(t) = [\, S(t) \ \ +/- \ \surd(S(t)^2 - \frac{4nS(t)\eta}{1-\eta})\ ] \ / \ 2n$$

In this case increase upon a positive experience will take place for $T_i^+(t)$ less than the smaller zero or higher than the larger zero, and not between the zeros. An equilibrium occurs upon a positive experience when $T_i^+(t) = 1$ or when equality holds:

$$n\, T_i^+(t)^2 - \ S(t) T_i^+(t) \ + \ \eta\, S(t) \, / (1 - \eta) = 0$$

This only can happen when the discriminant is not negative, in which case equilibria occur for $T_i^+(t)$ equal to one of the zeros.

**Case 2:** $E_i(t) \ = 0$
In this case the differential equation can be simplified to

$$\frac{dT_i^+(t)}{dt} = -\gamma * T_i^+(t)$$

So, in this case positive trust is decreasing or has in equilibrium with positive trust 0.

**Case 3:** $E_i(t) = -1$
In this case the differential equation can be simplified to

$$\frac{dT_i^+(t)}{dt} = 0$$

So, for this case always an equilibrium occurs in t for positive trust.

For negative trust, the situation is a mirror image of the case for positive trust, and by combining the positive and negative trust, the patterns for overall trust can be analyzed.

## 5 A Neural Model for Relative Trust

The model for trust dynamics from a neurological perspective is presented in this section. First, the background theory behind the model is presented. Thereafter, the model itself is explained in more detail. Simulation results are presented, and the resulting traces are formally analyzed to see whether they indeed show the appropriate patterns.

### 5.1 Background principles

Cognitive states of a person, such as sensory or other representations often induce emotions felt within this person, as described by neurologist Damasio (Damasio, 1999) and (Damasio, 2004); for example:

> 'Through either innate design or by learning, we react to most, perhaps all, objects with emotions, however weak, and subsequent feelings, however feeble.' (Damasio, 2004, p. 93)

In some more detail, emotion generation via a body loop roughly proceeds according to the following causal chain; see (Damasio, 1999) and (Damasio, 2004):

> cognitive state → preparation for bodily response → bodily response →
> sensing the bodily response → sensory representation of the bodily response → feeling

The body loop (or as if body loop) is extended to a recursive body loop (or recursive as if body loop) by assuming that the preparation of the bodily response is also affected by the state of feeling the emotion as an additional causal relation: feeling → preparation for the bodily response. Such recursiveness is also assumed by Damasio (Damasio, 2004), as he notices that what is felt by sensing is actually a body state which is an internal object, under control of the person:

> 'The brain has a direct means to respond to the object as feelings unfold because the object at the origin is inside the body, rather than external to it. (…) The object at the origin on the one hand, and the brain map of that object on the other, can influence each other in a sort of reverberative process that is not to be found, for example, in the perception of an external object.' ((Damasio, 2004) p. 91)

Within the model presented in this paper, both the bodily response and the feeling are assigned a level or gradation, expressed by a number. The causal cycle is modeled as a positive feedback loop, triggered by a mental state and converging to a certain level of feeling and body state.

Another neurological theory addressing the interaction between cognitive and affective aspects can be found in Damasio's Somatic Marker Hypothesis; cf. (Damasio, 1996), (Damasio, 1999), (Damasio, 2004) and (Bechara and Damasio, 2004). This is a theory on decision making which provides a central role to emotions felt. Within a given context, each represented decision option induces (via an emotional response) a feeling which is used to mark the option. For example, a strongly negative somatic marker linked to a particular option occurs as a strongly negative feeling for that option. Similarly, a positive somatic marker occurs as a positive feeling for that option. Damasio describes the use of somatic markers in the following way:

> 'the somatic marker (..) forces attention on the negative outcome to which a given action may lead, and functions as an automated alarm signal which says: beware of danger ahead if you choose the option which leads to this outcome. The signal may lead you to reject, *immediately*, the negative course of action and thus make you choose among other alternatives. (…) When a positive somatic marker is juxtaposed instead, it becomes a beacon of incentive..' ((Damasio, 1994) pp. 173-174)

Somatic markers may be innate, but may also by adaptive, related to experiences:

> 'Somatic markers are thus acquired through experience, under the control of an internal preference system and

under the influence of an external set of circumstances ...' ((Damasio, 1994) p. 179)

In the model introduced below, this adaptive aspect will be modeled as Hebbian learning; cf. (Hebb, 1949; Bi and Poo, 2001; Gerstner and Kistler 2002). Viewed informally, in the first place it results in a dynamical connection strength obtained as an accumulation of experiences over time (1). Secondly, in decision making this connection plays a crucial role as it determines the emotion felt for this option, which is used as a main decision criterion (2). As discussed in the introduction, these two properties (1) and (2) are considered two main functional, cognitive properties of a trust state. Therefore they give support to the assumption that the strength of this connection can be interpreted as a representation of the trust in the option considered.

## 5.2 Neural Model for Relative Trust

Informally described theories in scientific disciplines, for example, in biological or neurological contexts, often are formulated in terms of causal relationships or in terms of dynamical systems. To adequately formalize such a theory the hybrid dynamic modeling language LEADSTO has been developed that subsumes qualitative and quantitative causal relationships, and dynamical systems; cf. (Bosse, Jonker, Meij, and Treur 2007a). This language has been proven successful in a number of contexts, varying from biochemical processes that make up the dynamics of cell behavior (Jonker, Snoep, Treur, Westerhoff, and Wijngaards 2008) to neurological and cognitive processes (Bosse, Jonker, Los, Torre, and Treur, 2007b; Bosse, Jonker, and Treur, 2007c; Bosse, Jonker, and Treur, 2008). Within LEADSTO a

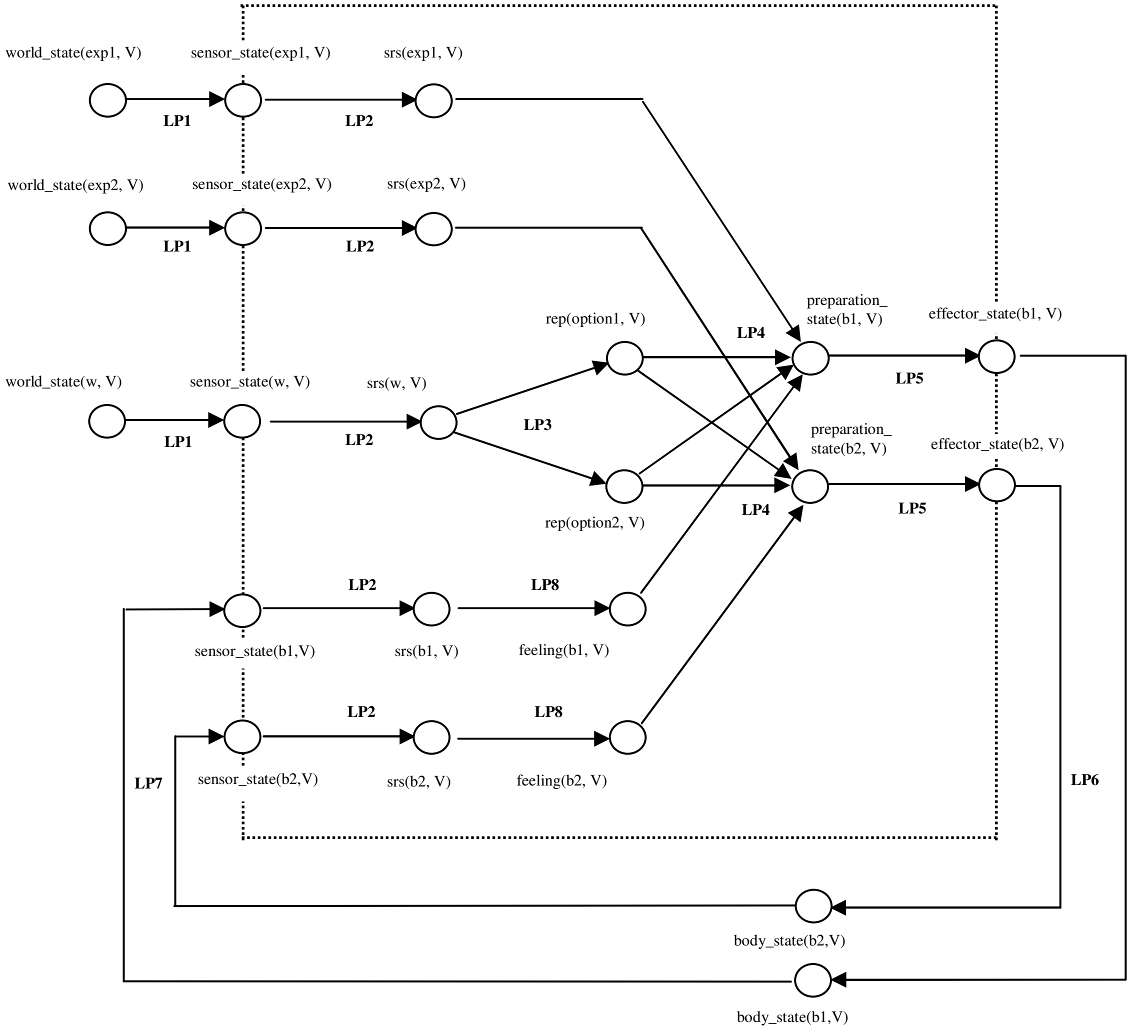


**Fig. 8.** Overview of the neural model for trust dynamics

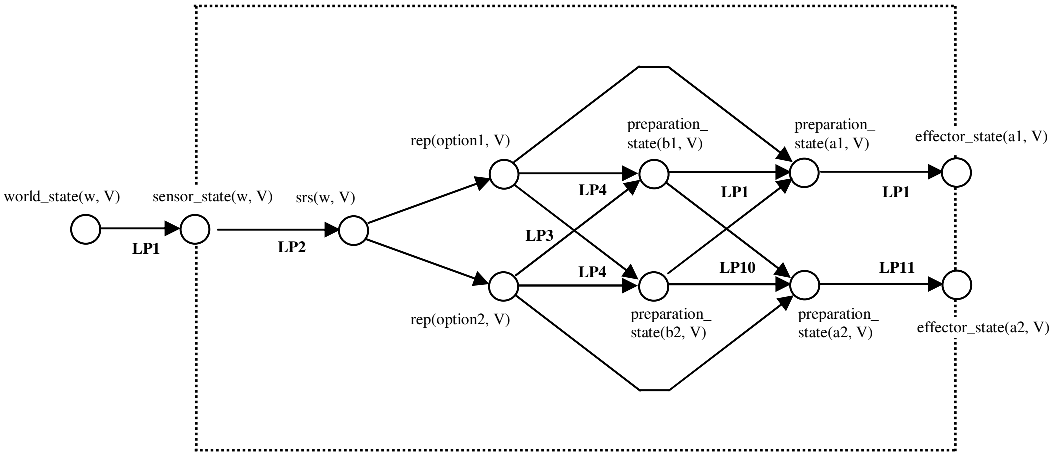


**Fig. 9.** Overview of the neural model for trust-based decision making

temporal relation a $\twoheadrightarrow$ b denotes that when a state property a occurs, then after a certain time delay (which for each relation instance can be specified as any positive real number), state property b will occur. In LEADSTO both logical and numerical calculations can be specified in an integrated manner; a dedicated software environment is available to support specification and simulation.

An overview of the model for how trust dynamics emerges from the experiences is depicted in Figure 8. How decisions are made, given these trust states is depicted in Figure 9. These pictures also show representations from the detailed specifications explained below. However, note that the precise numerical relations between the indicated variables V shown are not expressed in this picture, but in the detailed specifications of properties below, which are labeled by LP1 to LP11 as also shown in the pictures. The detailed specification (both informally and formally) of the model is presented below. Here capitals are used for (assumed universally quantified) variables. First the part is presented that describes the basic mechanisms to generate a belief state and the associated feeling. The first dynamic property addresses how properties of the world state can be sensed.

**LP1 Sensing a world state**

If world state property W occurs of strength *V*
then a sensor state for W of strength *V* will occur.
world_state(W, V) $\twoheadrightarrow$ sensor_state(W, V)

Note that this generic dynamic property is used for a specific world state, for experiences with the different options and for body states; to this end the variable W is instantiated respectively by w, exp1 and exp2, b1 and b2. From the sensor states, sensory representations are generated according to the dynamic property LP2. Note that also here for the example the variable P is instantiated as indicated.

**LP2 Generating a sensory representation for a sensed world or body state**

If a sensor state for world state or body state property P with level *V* occurs,
then a sensory representation for P with level *V* will occur.
sensor_state(P, V) $\twoheadrightarrow$ srs(P, V)

For a given world state representations for a number of options are activated:

**LP3 Generating an option for a sensory representation of a world state**

If a sensory representation for w with level *V* occurs
then a representation for optinon o with level *V* will occur
srs(w, V) $\twoheadrightarrow$ rep(o, V)

Dynamic property LP4 describes the emotional response to the person's mental state in the form of the preparation for a specific bodily reaction. Here the mental state comprises a number of cognitive and affective aspects: options activated, experienced results of options and feelings. This specifies part of the loop between feeling and body state. This property uses a combination model based on a function $g(\beta_1, \beta_2, V_1, V_2, V_3, \omega_1, \omega_2, \omega_3)$ including a threshold function. For example,

$$g(\beta_1, \beta_2, V_1, V_2, V_3, \omega_1, \omega_2, \omega_3) = th(\beta_1, \beta_2, \omega_1 V_1 + \omega_2 V_2 + \omega_3 V_3)$$

with $V_1$, $V_2$, $V_3$ activation levels and $\omega_1$, $\omega_2$, $\omega_3$ weights of the connections, and threshold function $th(\beta_1, \beta_2, V) = 1/(1+e^{-\beta_2(V-\beta_1)})$ with threshold $\beta_1$ and steepness $\beta_2$.

**LP4a From option activation and experience to preparation of a body state (non-competitive case)**

If option o with level $V_1$ occurs
and feeling the associated body state b has level $V_2$
and an experience for o occurs with level $V_3$
and the preparation state for b has level $V_4$
then a preparation state for body state b will occur with level $V_4 + \gamma(g(\beta_1, \beta_2, V_1, V_2, V_3, \omega_1, \omega_2, \omega_3)-V_4)\,\Delta t$.

rep(o, $V_1$) & feeling(b, $V_2$) & srs(exp, $V_3$) & preparation_state(b, $V_4$)
$\twoheadrightarrow$ preparation_state(b,
$V_4$+ $\gamma(g(\beta_1, \beta_2, V_1, V_2, V_3, \omega_1, \omega_2, \omega_3)-V_4)\,\Delta t$)

For the competitive case also the inhibiting cross connections from one represented option to the body state induced by another represented option are used. A function involving these cross connections was defined, for example

$$h(\beta_1, \beta_2, V_1, V_2, V_3, V_{21}, \omega_1, \omega_2, \omega_3, \omega_{21}) = th(\beta_1, \beta_2, \omega_1 V_1 + \omega_2 V_2 + \omega_3 V_3 - \omega_{21} V_{21})$$

for two considered options, with $\omega_{21}$ the weight of the suppressing connection from represented option 2 to the preparation state induced by option 1.

**LP4b From option activation and experience to preparation of a body state (competitive case)**

If option o1 with level $V_1$ occurs
and option o2 with level $V_{21}$ occurs
and feeling the associated body state b1 has level $V_2$
and an experience for o1 occurs with level $V_3$
and the preparation state for b1 has level $V_4$
then a preparation state for body state b1 will occur with level $V_4$+ $\gamma(h(\beta_1, \beta_2, V_1, V_2, V_3, V_{21}, \omega_1, \omega_2, \omega_3, \omega_{21})-V_4)\,\Delta t$.

rep(o1, $V_1$) & rep(o2, $V_{21}$) & feeling(b1, $V_2$) & srs(exp1, $V_3$) & preparation_state(b1, $V_4$)
$\twoheadrightarrow$ preparation_state(b1,
$V_4$+ $\gamma(h(\beta_1, \beta_2, V_1, V_2, V_3, V_{21}, \omega_1, \omega_2, \omega_3, \omega_{21})-V_4)\,\Delta t$)

Dynamic properties LP5, LP6, and LP7 together with LP2 describe the body loop.

**LP5 From preparation to effector state for body modification**

If preparation state for body state B occurs with level $V$,
then the effector state for body state B with level $V$ will occur.

preparation_state(B, V) $\twoheadrightarrow$ effector_state(B, V)

**LP6 From effector state to modified body state**

If the effector state for body state B with level $V$ occurs,
then the body state B with level $V$ will occur.

effector_state(B, V) $\twoheadrightarrow$ body_state(B, V)

**LP7 Sensing a body state**

If body state B with level $V$ occurs,
then this body state B with level $V$ will be sensed.

body_state(B, V) $\twoheadrightarrow$ sensor_state(B, V)

**LP8 From sensory representation of body state to feeling**

If a sensory representation for body state B with level $V$ occurs,
then B will be felt with level $V$.

srs(B, V) $\twoheadrightarrow$ feeling(B, V)

Alternatively, dynamic properties LP5 to LP7 can also be replaced by one dynamic property LP9 describing an as if body loop as follows.

**LP9 From preparation to sensed body state**

If preparation state for body state B occurs with level $V$,
then the effector state for body state B with level $V$ will occur.

preparation_state(B, V) $\twoheadrightarrow$ srs(B, V)

For the decision process on which option $O_i$ to choose, represented by action $A_i$, a winner-takes-it-all model is used based on the feeling levels associated to the options; for an overview, see Fig. 8.

This has been realised by combining the option representations $O_i$ with their related emotional responses $B_i$ in such a way that for each $i$ the level of the emotional response $B_i$ has a strongly positive effect on preparation of the action $A_i$ related to option $O_i$ itself, but a strongly suppressing effect on the preparations for actions $A_j$ related to the other options $O_j$ for $j \neq i$. As before, this is described by a function

$$h(\beta_1, \beta_2, V_1, \dots, V_m, U_1, \dots, U_m, \omega_{11}, \dots, \omega_{mm})$$

with $V_i$ levels for representations of options $O_i$ and $U_i$ levels of preparation states for body state $B_i$ related to options $O_i$ and $\omega_{ij}$ the strength of the connection between preparation states for body state $B_i$ and preparation states for action $A_j$.

**LP10 Decisions based on felt emotions induced by the options**

If options $O_i$ with levels $V_i$ occur,
and preparation states for body state $B_i$ related to options $O_i$ occur with level $U_i$,
and the preparation state for action $A_i$ for option $O_i$ has level $W_i$
then the preparation state for action $A_i$ for option $O_i$ will occur with level $W_i$ +
$\gamma(h(\beta_1, \beta_2, V_1, \dots, V_m, U_1, \dots, U_m, \omega_{11}, .. \omega_{mm}) - W_i)\,\Delta t$

rep($O_1$, $V_1$) & … & rep($O_m$, $V_m$) & preparation_state($B_1$, $U_1$) & … & preparation_state($B_m$, $U_m$) & preparation_state($A_i$, $W_i$)
$\twoheadrightarrow$ preparation_state($A_i$, $W_i$ +
$\gamma(h(\beta_1, \beta_2, V_1, \dots, V_m, U_1, \dots, U_m, \omega_{11}, .. \omega_{mm}) - W_i)\,\Delta t$)

**LP11 From preparation to effector state for an action**

If preparation state for action A occurs with level $V$,
then the effector state for action A with level $V$ will occur.

preparation_state(A, V) $\twoheadrightarrow$ effector_state(A, V)

**Hebbian Adaptation.** From a neurological perspective the strength of a connection from an option to an emotional response may depend on how experiences are felt emotionally, as neurons involved in the option, the preparation for the body state, and in the associated feeling will often be activated simultaneously. Therefore such a connection from option to emotional response may be strengthened based on a general Hebbian learning mechanism (Hebb, 1949; Bi and Poo 2001; Gerstner and Kistler 2002) that states that connections between neurons that are activated simultaneously are strengthened, similar to what has been proposed for the emergence of mirror neurons; e.g., (Keysers, and Perrett, 2004) and (Keysers, and Gazzola, 2009). This principle is applied to the strength $\omega_1$ of the connection from option 1 to the emotional response expressed by body state b1. The following learning rule takes into account a maximal connection strength *1*, a learning rate $\eta$, and an extinction rate $\zeta$.

**LP12 Hebbian learning rule for the connection from option to preparation**

```
If      the connection from option o1 to preparation of b1 has
            strength $\omega_1$
  and   the option o1 has strength $V_1$
  and   the preparation of b1 has strength $V_2$
  and   the learning rate from option o1 to preparation of b1
            is $\eta$
  and   the extinction rate from option o1 to preparation of b1
            is $\zeta$
then    after $\Delta t$ the connection from option o1 to preparation
            state b1 will have
            strength $\omega_1 + (\eta V_1 V_2 (1 - \omega_1) - \zeta\omega_1)\, \Delta t$.
```

```
has_connection_strength(rep(o1),preparation(b1),  ω1)  &
rep(o1, V1) & preparation(b1, V2) &
has_learning_rate(rep(o1), preparation(b1), η)  &
has_extinction_rate(rep(o1), preparation(b1), ζ)
→ has_connection_strength(rep(o1), preparation(b1),
        ω1 + (ηV1V2 (1 - ω1) - ζω1) Δt)
```

By this rule through their affective aspects, the experiences are accumulated in the connection strength from option o1 to preparation of body state b1, and thus serves as a representation of trust in this option o1. A similar Hebbian learning rule can be found in ((Gerstner and Kistler, 2002) p. 406).

# 6 Simulations for the Neural Model

The model described in Section 4 has been used to generate a number of simulation experiments for non-competitive and competitive cases (see Fig. 10 for some example results). To ease the comparison between these cases the same model parameter values were used for these examples (see Table 8). In Fig. 10a) example simulation results are shown for the non-competitive case. Here the subject is exposed to an information source that provides experience values *0* respectively *1* alternating periodically in a period of 200 time steps each. In this figure it can be observed that change in experience leads to changes in the connection strengths (representing trust) as well as the action effector states.

Furthermore, the decrease in the connection strengths representing trust due to a bad experience (0) takes longer than the increase due to a good experience (1), which can be explained by the higher value of the learning rate than of the extinction rate.

In Figures 10b), c) and d), the simulation results are shown for the competitive case with two competitive options having suppression weight *0.5* from option representation to preparation state and from preparation state to the action state. In this case the subject is exposed to two information sources that provides experience values *0* respectively *1* alternating periodically in a period of 200 time steps each, in a reverse cycle with respect to each other (see Figure 10b)). Here change in experience changes the connections representing trust as well as the action effector states. Moreover, in comparison to the non-competitive case, the learning is slow while decay is fast, which is due to the presence of competition. Finally Figure 10 shows that the connection strengths in the presented model exhibit the two fundamental functional properties of trust discussed in section 1, namely that trust is based on accumulation of experiences over time (see Figure 10c) and that trust states are used in decision making by choosing more trusted options above less trusted ones (see Figure 10d).

**Table 8.** Parameter values used in the example simulations

| Parameter | Value | Meaning |
|---|---|---|
| $\beta_1$ | 0.95 | threshold value for preparation state and action effector state |
| $\beta_2$ | 10,<br>100 | steepness value for preparation state, action effector state |
| $\gamma$ | 0.90 | activation change rate |
| $\eta$ | 0.80 | learning rate of connection from option representation to preparation |
| $\zeta$ | 0.10 | extinction rate of connection from option representation to preparation |
| $\Delta t$ | 0.90 | time step |
| $\omega_s$, $\omega_a$ | 0.50 | suppressing weight from option representation to preparation state and from preparation state to the action state (competitive case) |

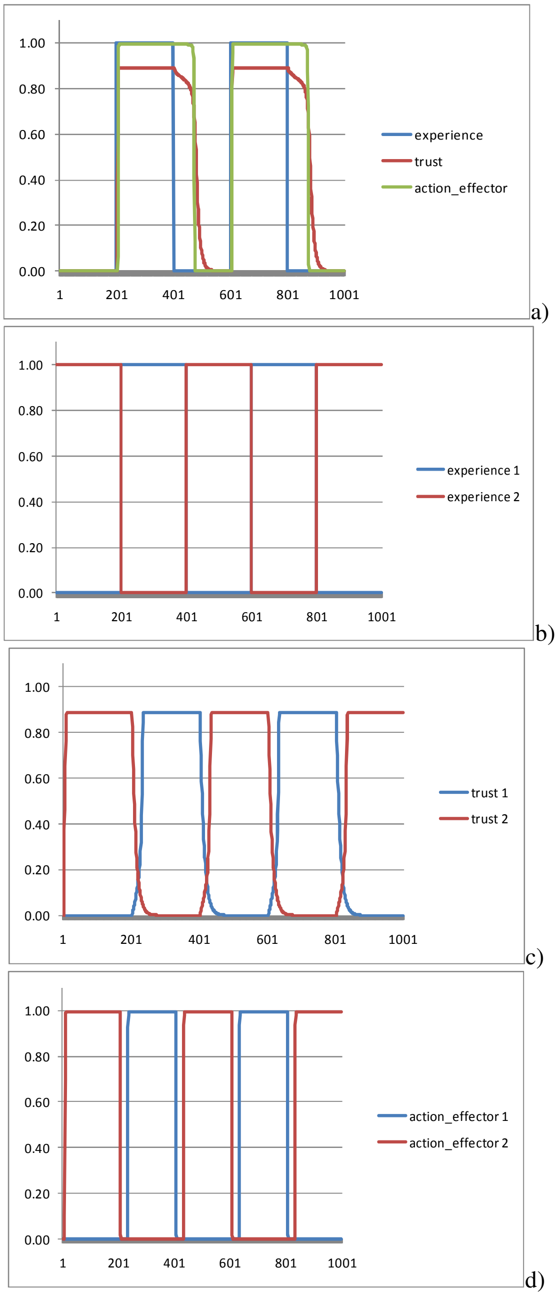


**Fig. 10.** Simulation Results: experience, connection representing trust, and action effector state for a non-competitive case in a) and a competitive case in b), c) and d) respectively

## 7 Analysis of the Neural Model

The two functional properties of trust states formulated in the introduction are: (1) A trust state result from accumulation of experiences over time, and (2) Trust states affect decision making by choosing more trusted above less trusted options. These properties characterize trust states from a functional, cognitive perspective. Therefore any model or computational or physical realization claimed to describe trust dynamics has to (at least) satisfy these properties. Such properties can be formalized in a temporal logical language, and can be automatically verified for the traces that have been generated using the proposed model. In this section this verification of properties is discussed. First, the language used to verify these properties is explained. Thereafter the properties and the results of the verification are discussed.

The verification of properties has been performed using a language called TTL (Temporal Trace Language), that features a dedicated editor and an automated checker; cf. (Bosse, Jonker, Meij, Sharpanskykh, and Treur, J. 2009). This predicate logical temporal language supports formal specification and analysis of dynamic properties, covering both qualitative and quantitative aspects. TTL is built on atoms referring to *states* of the world, *time points* and *traces*, i.e. trajectories of states over time. In addition, *dynamic properties* are temporal statements that can be formulated with respect to traces based on the state ontology Ont in the following manner. Given a trace $\gamma$ over state ontology Ont, the state in $\gamma$ at time point t is denoted by state($\gamma$, t). These states can be related to state properties via the infix predicate $\models$, where state($\gamma$, t) $\models$ p denotes that state property p holds in trace $\gamma$ at time t. Based on these statements, dynamic properties can be formulated using quantifiers over time and traces and the usual first-order logical connectives such as $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\forall$, $\exists$. For more details, see (Bosse, Jonker, Meij, Sharpanskykh, and Treur, J. 2009).

In order to be able to automatically verify the properties upon the simulation traces, they have been formalized. From the computational verification process it was found that indeed they are satisfied by the simulation traces of the model for which they were verified. The first functional property (1), specifying that a trust state accumulates experiences over time, is split up into a number of properties. First, two properties are specified which express trust accumulation for the non-competitive case, whereby the connections for the respective trustees are not influenced by experiences with competitors.

**P1.1 Connection strength increases with more positive experience (non-competitive case)**

If a sensor state indicates a particular value $V_1$ of an experience E, and E is an experience for trustee T, and the current strength of the connection for trustee T is

$V_2$, and $V_1$ is higher than $V_2$, then the connection strength will remain the same or increase.

```
∀γ:TRACE, t:TIME, E:EXPERIENCE, T:TRUSTEE,
V1,V2,V3:VALUE
    [ state(γ, t) |= sensor_state(E, V1) &
      state(γ, t) |= connection(T, V2) &
     state(γ, t) |= matches(E, T) & V1 > V2
       ⇒  ∃V3:VALUE [state(γ, t+1) |= connection(T, V3) &
          V3 ≥ V2 ] ]
```

**P1.2 Connection strength decreases with more negative experience (non-competitive case)**
If a sensor state indicates a particular value $V_1$ of an experience E, and E is an experience for trustee T, and the current strength of the connection for trustee T is $V_2$, and $V_1$ is lower than $V_2$, then the connection strength will remain the same or decrease.

```
∀γ:TRACE, t:TIME, E:EXPERIENCE, T:TRUSTEE,
V1,V2,V3:VALUE
[ state(γ, t) |= sensor_state(E, V1) &
  state(γ, t) |= connection(T, V2) &
  state(γ, t) |= matches(E, T) & V1 < V2
  ⇒ ∃V3:VALUE [ state(γ, t+1) |= connection(T, V3) &
     V3 ≤ V2 ] ]
```

Besides the non-competitive case, also properties have been specified for the competitive case. Hereby, the experiences with other competitive information sources are also taken into account.

**P2.1 Connection strength increases with more positive experience (competitive case)**
If a sensor state indicates a particular value $V_1$ of an experience E, and E is an experience for trustee T, and the current strength of the connection for trustee T is $V_2$, and $V_1$ is higher than $V_2$, and all other experiences are lower compared to $V_1$, then the connection strength will remain the same or increase.

```
∀γ:TRACE, t:TIME, E:EXPERIENCE, T:TRUSTEE,
V1,V2,V3:VALUE
[ state(γ, t) |= sensor_state(E, V1) &
   ∀E’ ≠ E [ ∃V’:VALUE state(γ, t) |= sensor_state(E’, V’) &
             V’ < V1 ] &
 state(γ, t) |= connection(T, V2) &
 state(γ, t) |= matches(E, T) & V1 > V2
⇒ ∃V3:VALUE [ state(γ, t+1) |= connection(T, V3) &
                    V3 ≥ V2 ] ]
```

**P2.2 Connection strength decreases with more negative experience (competitive case)**
If a sensor state indicates a particular value $V_1$ of an experience E, and E is an experience for trustee T, and the current strength of the connection for trustee T is $V_2$, and $V_1$ is lower than $V_2$, and all other experiences are higher compared to $V_1$, then the connection strength will remain the same or decrease.

```
∀γ:TRACE, t:TIME, E:EXPERIENCE, T:TRUSTEE,
V1,V2,V3:VALUE
[ state(γ, t) |= sensor_state(E, V1) &
   ∀E’ ≠ E [ ∃V’:VALUE state(γ, t) |= sensor_state(E’, V’) &
             V’  >V1 ] &
 state(γ, t) |= connection(T, V2) &
 state(γ, t) |= matches(E, T) & V1 < V2
  ⇒  ∃V3:VALUE [ state(γ, t+1) |= connection(T, V3) &
     V3 ≤ V2 ]]
```

Finally, property P3 is specified which compares different traces, as shown below.

**P3.1 Higher experiences lead to higher connection strengths (non-competitive case)**
If within one trace the experiences for a trustee are always higher compared to the experiences for a trustee in another trace, then in that trace the connection strengths will always be higher.

```
∀γ1, γ2:TRACE, E:EXPERIENCE, T:TRUSTEE
[ γ1≠ γ2 &  state(γ1, 0) |= matches(E, T) &
 ∀t:TIME [ ∃V1, V2:VALUE [
            state(γ1, t) |= sensor_state(E, V1) &
            state(γ2, t) |= sensor_state(E, V2) & V1>V2 ] ]
    ⇒   ∀t:TIME [ ∃V1, V2:VALUE [
                    state(γ1, t) |= connection(T, V1) &
                    state(γ2, t) |= connection(T, V2) &
                    V1>V2 ] ] ]
```

**P3.2 Higher experiences lead to higher connection strengths (competitive case)**
If within one trace the experiences for a trustee are always higher compared to the experiences for a trustee in another trace, and there are no other experiences with a higher value at that time point, then in that trace the connection strengths will always be higher.

```
∀γ1, γ2:TRACE, E:EXPERIENCE, T:TRUSTEE
[ γ1≠ γ2 &  state(γ1, 0) |= matches(E, T) &
  ∀t:TIME [ ∃V1, V2:VALUE [
     state(γ1, t) |= sensor_state(E, V1) &
     ∀E’ ≠ E [ ∃V’:VALUE
         state(γ1, t) |= sensor_state(E’, V’) & V’ ≤ V1 ] &
         state(γ2, t) |= sensor_state(E, V2) &
         ∀E” ≠ E [ ∃V”:VALUE
            state(γ2, t) |= sensor_state(E”, V”) & V” ≤ V2 ]   &
        V1>V2 ] ]
    ⇒  ∀t:TIME [ ∃V1, V2:VALUE [
           state(γ1, t) |= connection(T, V1) &
           state(γ2, t) |= connection(T, V2) & V1 > V2 ] ] ]
```

The formalization of the second functional property (2), i.e., trust states are used in decision making by choosing more trusted options above less trusted options, is expressed as follows.

**P4 The trustee with the strongest connection is selected**
If for trustee T the connection strength is the highest, then this trustee will be selected.

```
∀γ:TRACE, t1:TIME, T:TRUSTEE, V1:VALUE
 [ [ state(γ, t1) |= connection(T, V1)  &
    state(γ, t1) |= sensor_state(w, 1)  &
    ∀T’, V’ [ [ T’ ≠ T &
         state(γ, t1) |= connection(T’, V’) } ⇒ V’ < V1 ] ]
```

$\Rightarrow$ ∃t2:TIME < t1 + 10 [
    state(γ, t2) |= effector_state(T) ] ]

Note that in the property, the effector state merely has one argument, namely the trustee with the highest effector state value.

## 8. Comparing the Models by Mirroring

As mentioned before already, a direct comparison of the two models explained above is non-trivial as these models are described on a different level and each of the models include a specific set of parameters for cognitive and neurological characteristics of the person being modeled. The mapping between these parameters is difficult. As a consequence, relating the model from a formal perspective (i.e. relating the concepts of one model to the concepts of the other model) is not feasible. Of course, it is interesting to investigate whether the same patterns can result from the two different models to show that at least the same output can be provided given a similar input received by the model (and hence, show that the models are able to represent trust in a similar manner). In order to obtain this comparison, the models are compared in a more indirect way, by mutual *mirroring* them in each other.

The mirroring approach used to compare the two parameterized models for trust dynamics works as follows:

- Initially, for one of the models any set of values is assigned to its parameters
- Next, a number of scenarios are simulated based on this first model. These scenarios are carefully selected to allow for an investigation upon interesting experience sequences.
- The resulting simulation traces for the first model are approximated by the second model, based on automated parameter estimation.
- The error for the most optimal values for the parameters of the second model is considered as a comparison measure.

Parameter estimation can be performed according to different methods, for example, exhaustive search, bisection or simulated annealing (Hoogendoorn, Jaffry, and Treur, 2009). As the models considered here have only a small number of parameters exhaustive search is an adequate option. Using this method the entire attribute search space is explored to find the vector of parameter settings with maximum accuracy. This method guarantees the optimal solution, described as follows:

```
for each: observed behavior $B$
    for each: vector of parameter value settings $P$
        calculate the accuracy of $P$
    end for
    output: the vector of parameter settings with maximal accuracy
end for
```

In the above algorithm, calculation of the accuracy of a vector of parameter setting P entails that agent predicts the information source to be requested and observes the actual human request. It then uses the equation for calculating the accuracy described before. Here if $p$ parameters are to be estimated with precision $q$ (i.e., grain size $10^{-q}$), the number of options is $n$, and $m$ the number of observed outcomes (i.e., time points), then the worst case complexity of the method can be expressed as $O\ ((10)^{pq}\ nm^2)$, which is exponential in number of parameters and precision. In particular, when $p=3$ (i.e., the parameters $\beta$, $\gamma$, and $\eta$), $q=2$ (i.e., grain size $0.01$), $n=3$ and $m=100$, then the complexity will result in $3 \times 10^{10}$ steps.

A number of experiments were performed using the mutual mirroring approach described in section 5 to compare the two parameterized models for trust dynamics. Experiments were set up according to two cases:

1. Two competitive options provide experiences *deterministically*, with a constant positive, respectively negative experience, alternating periodically in a period of 50 time steps each (see Figure 11).
2. Two options provide experiences with a certain *probability* of positivity, again in an alternating period of 50 time steps each.

The first case of experiments was designed to compare the behavior of the models for different parameters under the same deterministic experiences while the second case is used to compare the behavior of the models for the (more realistic) case of probabilistic experience sequences. The general configurations of the experiment that are kept constant for all experiments are shown in Table 9.

Three experiments were performed for each case: after some parameter values assigned to the cognitive model, its behavior was approximated by the neural model, using the mirroring approach based on the automatic parameter estimation technique described above. The best approximating realization of the neural model was used again to approximate the cognitive model using the same mirroring approach. This second approximation was performed to minimize uni-directionality of the mirroring approach that might bias the results largely if performed from only one model to another and not the other way around.

**Table 9.** General Experimental Configuration

| Parameter | Neural Model | Cognitive Model |
|---|---|---|
| Number of competitive options | 2 | 2 |
| Time step (difference equations) | 0.1 | 0.1 |
| Number of time steps | 500 | 500 |
| Initial trust values of option 1 and option 2 | 0.5, 0.5 | 0, 0 |
| Strength of connection from option to emotional response ($\omega_1$) | 0.5 | not applicable |
| Strength of connection between preparation state of body and preparation state of action ($\omega_{ij}$) | 0.5 | not applicable |
| Strength of connection between feeling and preparation of body state | 0.25 | not applicable |
| Value of the world state | 1 | not applicable |
| Grain size in parameter estimation | 0.05 | 0.01 |

An instance of a parameterized model can uniquely be represented by a tuple containing the values of its parameters. Here the cognitive and neural models described in section 2 and 3 are represented by tuples ($\gamma$, $\beta$, $\eta$) and ($\sigma$, $\tau$, $\gamma$, $\eta$, $\zeta$) respectively. For the sake of simplicity, a few parameters of the neural model, namely $\omega_{1,}$ $\omega_{12}$ and $\omega_{21}$, were considered fixed with value *0.5*, and were not included in model representation tuple. Furthermore, the initial trust values of both models are assumed neutral (*0.0* and *0.5* for cognitive and neural model resp.), see Table 9.

### Case 1

In this case the behavior of the models was compared using the experiences that were provided deterministically with positive respectively negative, alternating periodically in a period of 50 time steps each (see Figure 11). Here three different experiments were performed, where the parameters of cognitive model are assigned with some initial values and then its behavior is approximated by the neural model. The best approximation of the neural model against the initially set cognitive model was reused to find the best matching cognitive model.

Results of the approximated models and errors are shown in Table 10 while the graphs of the trust dynamics are presented in Figure 12. Note that for the sake of ease of comparison and calculation of standard error the trust values of cognitive model are projected from the interval [*-1, 1*] to [*0, 1*] (see Figure 12).

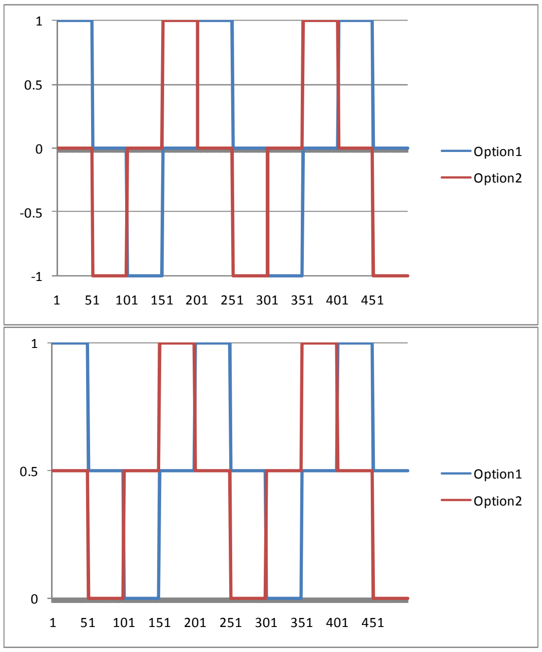


**Fig. 11.** a) Experience sequence for cognitive model, b) Experience sequence for neural model

In Table 10, the comparison error $\varepsilon$ is the average of the root mean squared error of trust of all options, as defined by the following formula,

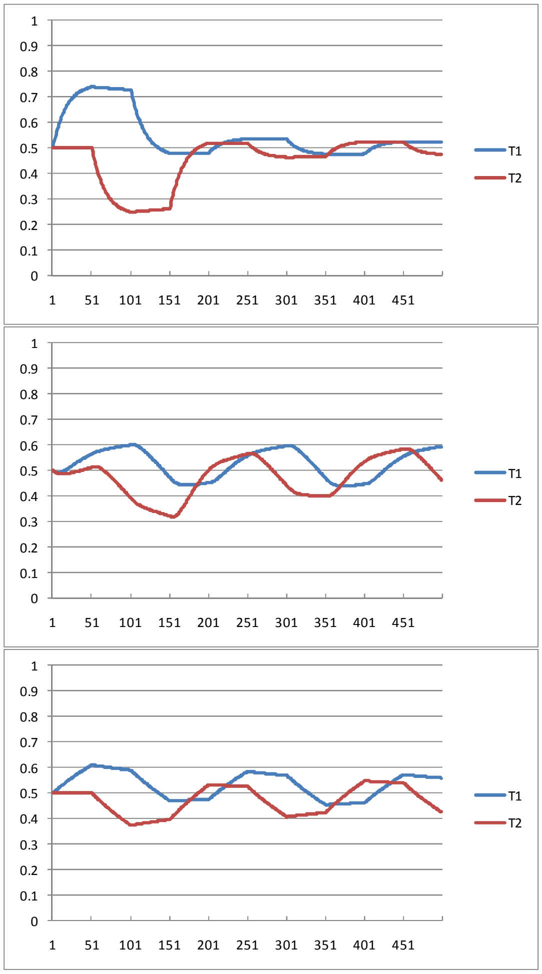
$$\varepsilon = \frac{1}{n} * \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} (T(j)_{1i} - T(j)_{2i})^2}$$

In the above formulation, $n$ is the number of options, $m$ is the number of time steps while $T(j)_{1i}$ and $T(j)_{2i}$ represent trust value of option $i$ at time point $j$ for each model, respectively.

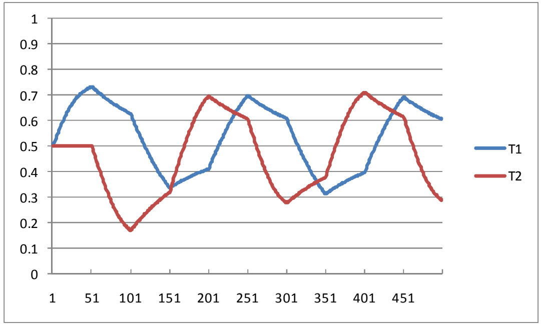
**Table 10.** Results of Case 1.

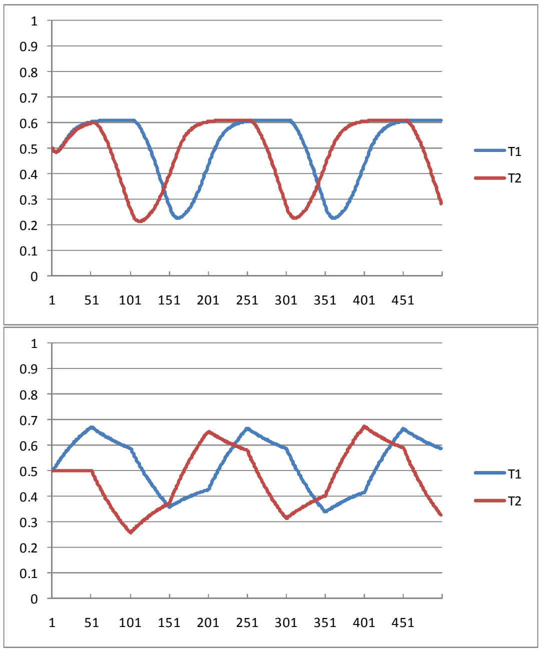
| Exp. | Initial Model | Approximating Model using the mirroring approach | Comparison Error ($\varepsilon$) |
|---|---|---|---|
| 1 | CM(*0.99, 0.75, 0.75*) | NM(*0.55, 10, 0.15, 0.90, 0.50*) | *0.074050* |
| | NM(*0.55, 10,0 .15, 0.90, 0.50*) | CM(*0.96, 0.20, 0.53*) | *0.034140* |
| 2 | CM(*0.88, 0.99, 0.33*) | NM(*0.35, 10, 0.60, 0.95, 0.60*) | *0.071900* |
| | NM(*0.35, 10, 0.60, 0.95, 0.60*) | CM(*0.87, 0.36, 0.53*) | *0.059928* |
| 3 | CM(*0.75, 0.75, 0.75*) | NM(*0.30, 10, 0.95, 0.90, 0.60*) | *0.138985* |
| | NM(*0.30, 10, 0.95, 0.90, 0.60*) | CM(*0.83, 0.37, 0.55*) | *0.075991* |

In Table 10 for experiment 1 initially the cognitive model was set with parameters (0.99, 0.75, 0.75) which was then approximated by the neural model. The best approximation of the neural model was found to be (0.55, 10, 0.15, 0.90, 0.50) with an approximate average of root mean squared error of all options $\varepsilon$ value 0.074050. Then this setting of neural model was used to approximate cognitive model producing best approximate with parameter values (0.96, 0.20, 0.53) producing $\varepsilon$ = 0.034140. Similarly the results of other two experiments can be read in Table 10.

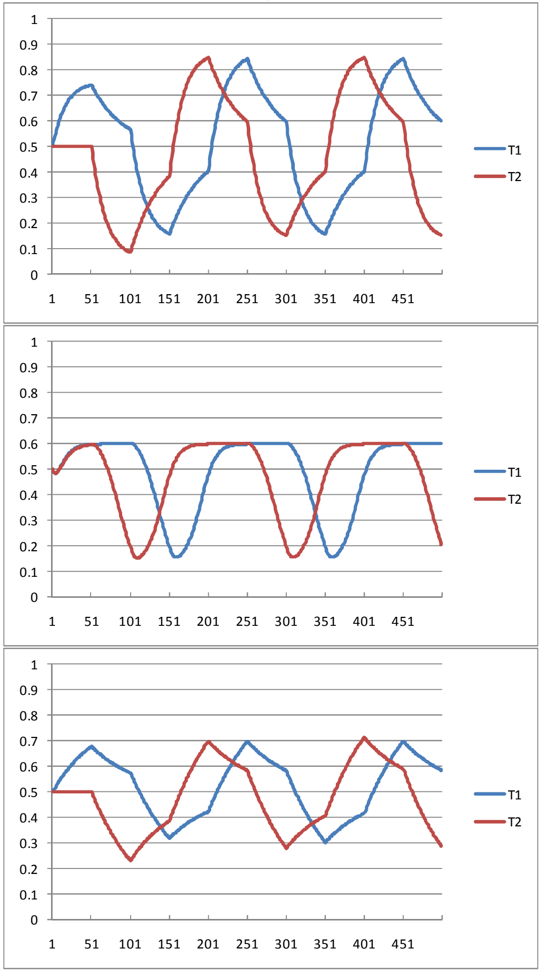


a)





b)



c)

**Fig. 12.** Dynamics of the Trust in Case 1 a) Experiment 1, b) Experiment 2, c) Experiment 3

Figure 12 represents the dynamics of the trust in the two options over time for the deterministic case. The horizontal axis represent time step while vertical axis represent the value of trust. The graphs for each experiment are represented as set of three figures, where the first figure shows the dynamics of the trust of both options by the cognitive model with an initial setting as described in the second column of the first row of each experiment of Table 10. The second figure shows the traces of the dynamics of trust by the neural model as described in the third column of the first row of each experiment of Table 10. Finally the third figure shows the approximation of the cognitive model by the neural model, where the neural model is described in the second column of the second row of each experiment of Table 10 (which is similar to third column of the first row of each experiment), and the approximated cognitive model is presented in the third column of the second row of each experiment. From Table 10 and Figure 12 it can be observed that the mirroring approach based on automatic parameter estimation when used in bidirectional way gives a better realization of both models in each other, resulting in a smaller comparison error and better curve fit.

### Case 2

In the second case the behavior of the models was compared when experiences are provided with a certain probability of positivity, again in an alternating period of 50 time steps each. Also here three different experiments were performed, where the parameters of the cognitive model were assigned with some initial values and then its behavior was approximated by the neural model. The best approximation of the neural model against initially set cognitive model was reused to find the best matching cognitive model. In experiment 1, 2 and 3 the option 1 and option 2 give positive experiences with (100, 0), (75, 25) and (50, 50) percent of probability, respectively. Results of approximated models and errors for this case are shown in Table 11while the graphs of trust dynamics are presented in Figure 13. Note that for the sake of ease of comparison and calculation of the standard error, again the trust values of the cognitive model are projected from the interval [*-1, 1*] to [*0, 1*] (see Figure 13). In Table 11 for experiment 1 initially the cognitive model was set with parameters (0.99, 0.75, 0.75) which was then approximated by the neural model.

**Table 11.** Results of Case 2.

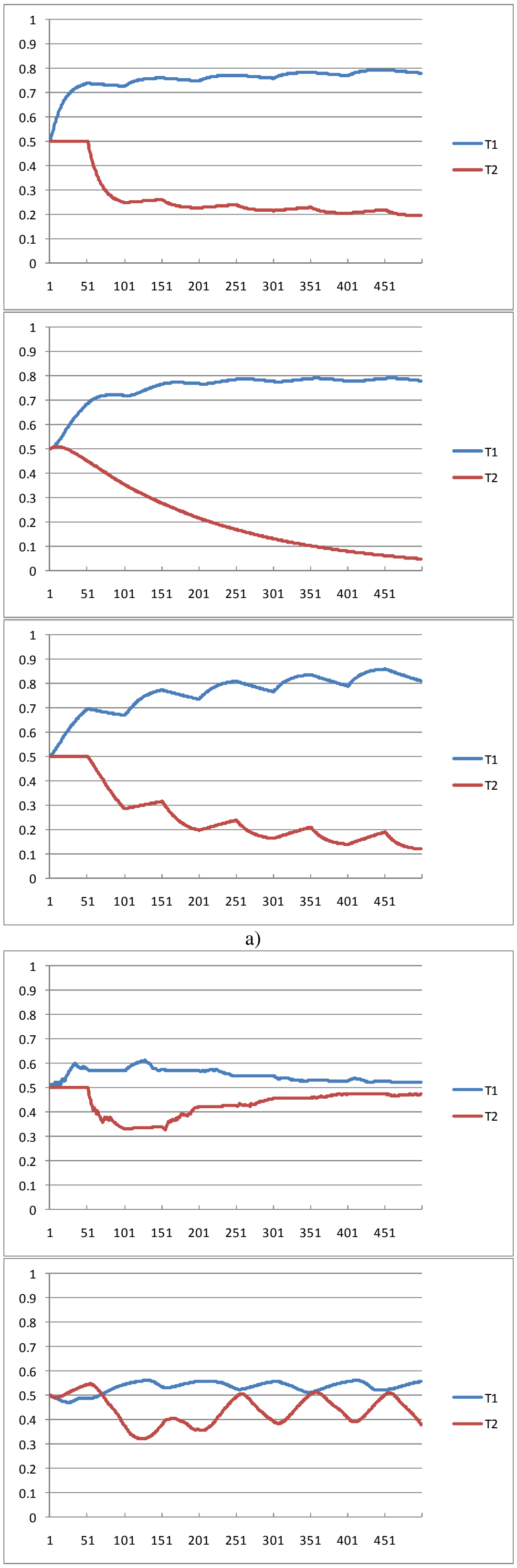
| Exp. | Initial Model | Approximating Model using the mirroring approach | Error ($\varepsilon$) |
|---|---|---|---|
| 1 | CM(0.99, 0.75, 0.75) | NM(0.85, 10, 0.95, 0.20, 0.05) | 0.061168 |
| | NM(0.85, 10, 0.95, 0.20, 0.05) | CM(0.97, 0.99, 0.18) | 0.045562 |
| 2 | CM(0.99, 0.75, 0.75) | NM(0.40, 20, 0.90, 0.20, 0.15) | 0.044144 |
| | NM(0.40, 20, 0.90, 0.20, 0.15) | CM(0.83, 0.05, 0.99) | 0.039939 |
| 3 | CM(0.99, 0.75, 0.75) | NM(0.10, 20, 0.45, 0.10, 0.10) | 0.011799 |
| | NM(0.10, 20, 0.45, 0.10, 0.10) | CM(0.99, 0.50, 0.99) | 0.011420 |

The best approximation of the neural model was found to be (0.85, 10, 0.95, 020, 0.05) with an approximate average of root mean squared error of all options $\varepsilon$ of value 0.061168. Then this setting of neural model was used to approximate cognitive model producing best approximate with parameter values (0.97, 0.99, 0.18) and $\varepsilon$ 0.034140. Similarly the results of other two experiments could also be read in Table 11.

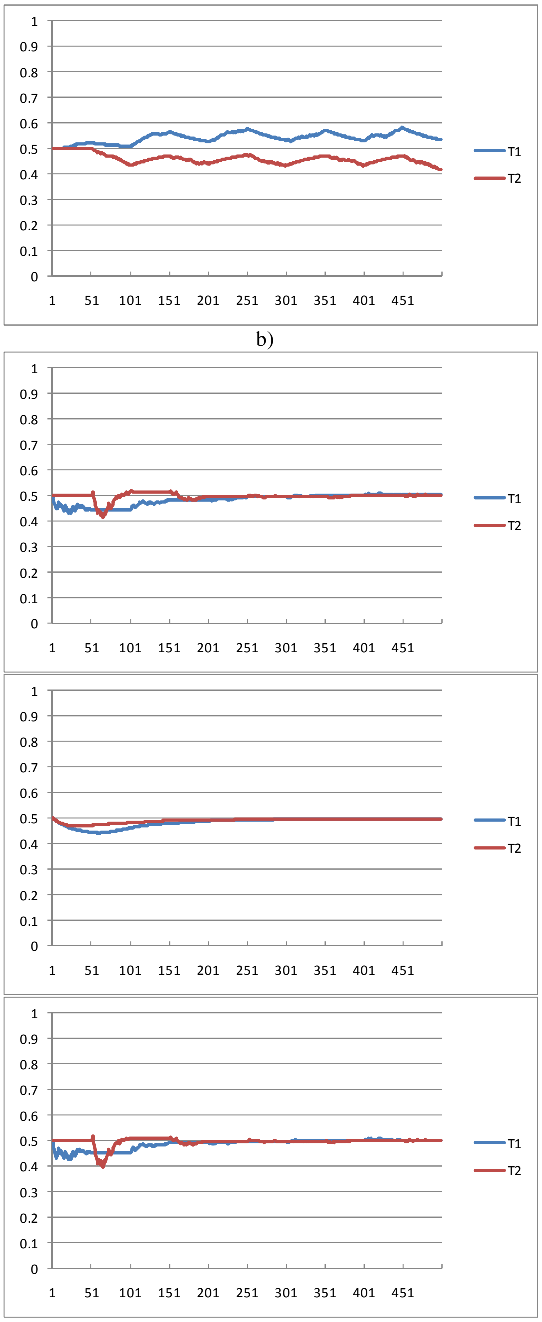
Fig. 13 represents the dynamics of the trust in the two options over time for the probabilistic case. The horizontal axis represents time while the vertical axis represents the values of trust. Here also the graphs of each experiment are represented as set of three figures, where the first figure shows the dynamics of the trust in both options by the cognitive model with an initial setting as described in the second column of the first row of each experiment of Table 11.

The second figure shows the traces of the dynamics of trust by the neural model as described in the third column of the first row of each experiment of Table 11. Finally, the third figure is the approximated cognitive model by the neural model, where the neural model is described in the second column of the second row of each experiment of Table 11 (which is similar to third column of the first row of each experiment), and the approximated model is presented in the third column of the second row of each experiment.

As already noticed in case 1, also here it can be observed that the mirroring approach based on automatic parameter estimation when used in bidirectional way gives a better realization of both models in each other, resulting smaller comparison error and a better curve fit. Furthermore, it can also be noted that as the uncertainty in the options behavior increases, both models show more similar trust dynamics producing lower error value in comparison.



a)



b)

c)

**Fig. 13.** Dynamics of the Trust in Case 2, a) Experiment 1, b) Experiment 2, c) Experiment 3

## 9 Discussion

This paper has introduced a cognitive and a neural model for relative trust in competitive trustees. Within these models parameters have been introduced to tailor it towards a particular human. Simulation experiments have been run, and formal verification techniques have been applied to show that models indeed exhibit desired patterns. A variety of trust models have been proposed in the literature (Falcone and Castelfranchi, 2004; Jonker and Treur, 1999). These trust models attempt to determine the level of trust in certain agents based upon experiences. They do however not take into account the notion of relativeness of this trust. Models have been proposed for relative trust as well. In (Beth, Borcherding and Klein, 1994) a model is presented that allows an agent to combine multiple sources for deriving a trust value. This notion of relativeness differs from the notion used in this paper. (Kluwer and Waaler, 2006) extends an existing trust model of (Jones, 2002) with the notion of relative trust. They take as a basis certain trust values determined by the model (Jones, 2002), and compare these values in order to make statements about different trust values for different agents. In determining the trust itself, they do not incorporate the experiences with other agents that can perform similar tasks, which is done in this paper. In (Maanen and Dongen, 2005) a trust model is utilized to allocate decision support tasks. In the model, relative trust is addressed as well but again not incorporated in the calculation of the trust value itself.

The proposed neural model incorporates the reciprocal interaction between cognitive and affective aspects based on neurological theories that address the role of emotions and feelings. The model describes more specifically how considered options and experiences generate an emotional response that is felt. For feeling the emotion, based on elements taken from (Damasio, 1999), (Damasio, 2003) and (Bosse, Jonker, and Treur, 2008), a converging recursive body loop is included in the model. An adaptation process based on Hebbian learning (Hebb, 1949, Bi and Poo, 2001, Gerstner and Kistler, 2002), was incorporated, inspired by the Somatic Marker Hypothesis described in (Damasio, 1994), (Damasio, 1996) and (Bechara and Damasio, 2004), and as also has been proposed for the functioning of mirror neurons; e.g., (Keysers and Perrett, 2004) and (Keysers and Gazzola, 2009). The idea of somatic marking is very general and functions as a kind of integrating factor in practically all mental processes, in particular in those in which affective and cognitive states interact in an adaptive manner. Therefore it is a quite useful modelling concept with a wide applicability; for a different application of this concept, in particular to model interacting cognitive and affective aspects of desires, see (Bosse, Hoogendoorn, Memon, Treur, and Umair, 2010).

The model was specified in the hybrid dynamic modeling language LEADSTO, and simulations were performed in its software environment; cf. (Bosse, Jonker, Meij, and Treur, 2007a). It has been shown that within the model the strength of the connection between a considered option and the emotional response induced by it, satisfies two properties that are considered as two central functional properties defining the causal or functional role of a trust state as a cognitive state (Jonker and Treur, 2003):

(1) it results from accumulation of experiences, and
(2) it affects deciding for the option.

This provides support for the assumption that the strength of this connection can be interpreted as a representation of the trust level in the considered option. Models of neural processes can be specified at different levels of abstraction. The model presented here can be viewed as a model at a higher abstraction level, compared to more detailed models that take into account more fine-grained descriptions of neurons and their biochemical and/or spiking patterns. States in the presented model can be viewed as abstract descriptions of states of neurons or as representing states of groups of neurons. An advantage of a more abstract representation is that such a model is less complex and therefore may be less difficult to handle, while it can still be shown that the model is able to express the essential dynamics of trust.

In addition, in the paper the proposed cognitive and neural model of trust have been compared. As the parameter sets for both models are different, the comparison involved mutual estimation of parameter values by which the models were mirrored into each other in the following manner. Initially, for one of the models any set of values was assigned to the parameters of the model, after which a number of scenarios were simulated based on this first model. Next, the resulting simulation traces for this first model were approximated by the second model, based on automated parameter estimation. The error for the most optimal values for the parameters of the second model was considered as a comparison measure. It turned out that approximations could be obtained with error margins of up to about 10%. Here the results for the (more realistic) case of probabilistic experience sequences show a better approximation than for the deterministic case. This can be considered a positive result, as the two models have been designed in an independent manner, using totally different concepts and techniques. In particular, it shows that the cognitive model, which was designed first, without taking into account neurological knowledge, can still be grounded in a neurological context, which is a nontrivial result.

Summarising, the claims of the paper that have been justified positively are that:

(1) appropriate patterns of trust dynamics are generated by the models,
(2) the models are based upon relevant theories from the respective domains, and
(3) they can be compared to each other in a reasonably accurate manner.

For future work, an interesting option is to see how well the parameters of the models can be derived by a personal assistant (based upon the requests provided as output by the human). Also how such a more abstract models like neural model of relative trust can be related to more detailed models, and in how far patterns observed in more specific models also are represented in such a more abstract model. Furthermore, part of future work is to validate the model based upon empirical data obtained from experiments.

## References

1. Aarts, E., Harwig, R., and Schuurmans, M. (2001). Ambient Intelligence. In P. Denning (ed.), *The Invisible Future* (pp. 235-250). New York:McGraw Hill. .
2. Aarts, E., Collier, R., van Loenen, E., & Ruyter, B. de (eds.) (2003). *Ambient Intelligence.* (Vol. 2875. pp. 432): Springer Verlag.
3. Bechara, A., & Damasio, A. (2004). The Somatic Marker Hypothesis: a neural theory of economic decision. *Games and Economic Behavior, 52*, 336-372.
4. Beth, T., Borcherding, M., & Klein, B. (1994) "Valuation of trust in open networks". Paper presented at 3rd European Symposium on Research in Computer Security, (ESORICS).
5. Bi, G.Q., & Poo, M.M. (2001) Synaptic Modifications by Correlated Activity: Hebb's Postulate Revisited. *Annual Review Neuroscience, 24*, 139-166.
6. Bosse, T., Hoogendoorn, M., Memon, Z.A., Treur, J., and Umair, M., (2010). An Adaptive Model for Dynamics of Desiring and Feeling based on Hebbian Learning. In: Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., and Huang, J. (eds.), *Proceedings of the Second International Conference on Brain Informatics, BI'10*. Lecture Notes in Artificial Intelligence, vol. 6334, Springer Verlag, 2010, pp. 14-28.
7. Bosse, T., Jonker, C.M., Meij, L. van der, & Treur, J. (2007a). A Language and Environment for Analysis of Dynamics by Simulation. *International Journal of Artificial Intelligence Tools*, *16*, 435-464.
8. Bosse, T., Jonker, C.M., Los, S.A., Torre, L. van der, & Treur, J. (2007b). Formal Analysis of Trace Conditioning. *Cognitive Systems Research Journal*, *8*, 36-47.
9. Bosse, T., Jonker, C.M., & Treur, J. (2007c). Simulation and Analysis of Adaptive Agents: an Integrative Modelling Approach. *Advances in Complex Systems Journal*, *10*, 335-357.
10. Bosse, T., Jonker, C.M., & Treur, J. (2008). Formalisation of Damasio's Theory of Emotion, Feeling and Core Consciousness. *Consciousness and Cognition Journal*, *17*, 94–113.
11. Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A., & Treur, J. (2009). Specification and Verification of Dynamics in Agent Models. *International Journal of Cooperative Information Systems*, *18*, 167-193.
12. Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Papermac.
13. Damasio, A. (1996). The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex. *Philosophical Transactions of the Royal Society: Biological Sciences*, *351*, 1413-1420
14. Damasio, A. (1999). *The Feeling of What Happens. Body and Emotion in the Making of Consciousness.* New York: Harcourt Brace.
15. Damasio, A. (2003). *Looking for Spinoza*. London: Vintage books.
16. Edward H. Shortliffe, & Bruce G. Buchanan (1975), A model of inexact reasoning in medicine. *Mathematical Biosciences, 23(3-4)*, 351-379
17. Eich, E., Kihlstrom, J.F., Bower, G.H., Forgas, J.P., & Niedenthal, P.M. (2000). *Cognition and Emotion*. New York: Oxford University Press.
18. Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself. Paper presented at 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004).
19. Forgas, J.P., Laham, S.M., & Vargas, P.T. (2005). Mood effects on eyewitness memory: Affective influences on susceptibility to misinformation. *Journal of Experimental Social Psychology*, *41*, 574–588.
20. Forgas, J.P., Goldenberg, L., & Unkelbach, C. (2009). Can bad weather improve your memory? An unobtrusive field study of natural mood effects on real-life memory. *Journal of Experimental Social Psychology*, *45*, 254–257.
21. Hebb, D. (1949). *The Organisation of Behavior*. New York: Wiley.
22. Hoogendoorn, M., Jaffry, S.W., & Treur, J., (2009). An Adaptive Agent Model Estimating Human Trust in Information Sources. Paper presented at 9th IEEE/WIC/ACM International Conference on Intelligent Agent Technology,(IAT'09).
23. Gerstner, W., & Kistler, W.M. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics*, *87*, 404–415
24. Jones, A. (2002). On the concept of trust, *Decision Support Systems*, *33*, 225-232.
25. Jonker, C.M., Snoep, J.L., Treur, J., Westerhoff, H.V., & Wijngaards, W.C.A. (2008). BDI-Modelling of Complex Intracellular Dynamics. *Journal of Theoretical Biology*, *251*, 1–23.
26. Jonker, C.M., & Treur, J., (2003). A Temporal-Interactivist Perspective on the Dynamics of Mental States. *Cognitive Systems Research Journal, 4*, 137-155.
27. Jonker, C.M., & Treur, J., (1999). Formal Analysis of Models for the Dynamics of Trust based on

Experiences. In F.J. Garijo, M. Boman (eds.), *Multi-Agent System Engineering, Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99.* LNAI vol. 1647, (pp. 221-232): Springer Verlag / Berlin.

28. Keysers, C., & Perrett, D.I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences*, *8*, 501-507.
29. Keysers, C., & Gazzola, V., (2009). Unifying Social Cognition. In: Pineda, J.A. (ed.), *Mirror Neuron Systems: the Role of Mirroring Processes in Social Cognition* (pp. 2-28). Humana Press/Springer Science.
30. Kluwer, J. & Waaler, A. (2006). Relative Trustworthiness, In Dimitrakos, T. *et al.* (eds), *Proceedings of Workshop on Formal Aspects of Security and Trust*, (pp.158-170).
31. Luger, G.F., & Stubblefield, W.A., (1998). Artificial Intelligence: Structures and Strategies for Complex Problem Solving. (4th ed.). Addison-Wesley.
32. Maanen, P.-P. van, & Dongen, K. van. (2005). Towards Task Allocation Decision Support by means of Cognitive Modeling of Trust, In C. Castelfranchi, S. Barber, J. Sabater, and M. Singh (Eds.), *Proceedings of the Eighth International Workshop on Trust in Agent Societies*, (pp. 168-77).
33. Marx, M., and Treur, J., (2001). Trust Dynamics Formalised in Temporal Logic. In L. Chen, Y. Zhuo(eds.), In *Proceeding. of the Third International Conference on Cognitive Science*, (pp. 359-363): USTC Press / Beijing.
34. Niedenthal, P.M. (2007). Embodying Emotion. *Science, 316*, 1002-1005.
35. Riva, G., F. Vatalaro, F. Davide, & M. Alcañiz. (2005). *Ambient Intelligence*. IOS Press.
36. Schooler, J.W., & Eich, E. (2000). Memory for Emotional Events. In: E. Tulving, F.I.M. Craik (eds.), *The Oxford Handbook of Memory* (pp. 379-394). Oxford University Press.
37. Winkielman, P., Niedenthal, P.M., & Oberman, L.M. (2009). Embodied Perspective on Emotion-Cognition Interactions. In: Pineda, J.A. (ed.), *Mirror Neuron Systems: the Role of Mirroring Processes in Social Cognition* (235-257). Humana Press/Springer Science.