Modeling and Validation of Biased Human Trust

Mark Hoogendoorn[1], S. Waqar Jaffry[1], and Peter-Paul van Maanen[1,2], and Jan Treur[1]
[1]VU University Amsterdam, Department of Artificial Intelligence,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
{mhoogen, swjaffry, treur}@cs.vu.nl
[2]TNO Human Factors, Department of Cognitive Systems Engineering,
P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
peter-paul.vanmaanen@tno.nl

***Abstract*— When considering intelligent agents that interact with humans, having an idea of the trust levels of the human, for example in other agents or services, can be of great importance. Most models of human trust that exist, are based on some rationality assumption, and biased behavior is not represented, whereas a vast literature in Cognitive and Social Sciences indicates that humans often exhibit non-rational, biased behavior with respect to trust. This paper reports how some variations of biased human trust models have been designed, analyzed and validated against empirical data. The results show that such biased trust models are able to predict human trust significantly better.**

Keywords: human trust, biases, modeling, validation

I. INTRODUCTION *(HEADING 1)*

Within multi-agent systems, trust is an essential concept that is usually considered as a means to aggregate direct or indirect experiences with particular issues, such as other agents or services; e.g., [13; 14; 15]. Such a trust value can for instance be taken into account when deciding on cooperation with other agents. A variety of computational trust models have been developed that express the formation of trust; see e.g. [10; 11] for an overview. Some of these trust models are intended to describe human trust; see e.g. [4; 9]. Human trust does not necessarily follow some optimality or rationality criterion in formation of trust, whereas trust models often have been designed with such a criterion in mind. Especially when developing an agent that interacts with humans, providing the agent with a good insight into the trust of humans can be beneficial and even necessary. Examples where such an insight can be useful include personal assistant agents that prepare the usage of certain outside sources whereby the source can be taken that is most trusted by the human. The agent could also maintain a model of the trust level of the human in the agent itself, whereby the behavior of the agent can be dependent on this estimated trust level; e.g., when the human has a low trust value in the agent, the agent could adopt a different strategy of communication.

When considering the literature on human trust characteristics, a variety of authors have shown that humans frequently show biases in their trust behavior. Among other factors, such biased behavior is very dependent on the culture. For example, [16] show that in collectivistic cultures humans tend to have a bias towards trusting members of the same group, whereas they are more negative towards people from outside the group. In [7] a comparison is made between individualistic and collectivistic cultures, and indeed it was shown that persons within an individualistic society tend to be less negatively biased towards persons from outside their group. Other authors also emphasize the existence of such a bias in general, e.g. [12]. If the objective of a computational model of trust is to create a model that represents human trust in a natural and accurate manner, such biases need to be taken into account in the model. In the models that have been proposed for human trust (e.g. [4; 9]) these biases are however not considered.

In this paper, a number of possibilities to model biased human trust in a computational manner are discussed and evaluated. In order to evaluate the newly developed biased trust models, and to show that they achieve better results in predicting human trust than nonbiased trust models, the models have been validated on empirical data obtained from experiments with humans. In the validation experiment, the humans were asked to perform a classification task whereby they received advice from several sources. In order to make the validation possible, the parameters of the model are tuned on an initial dataset of observed human behavior, and then used to predict the trust-based behavior in a second dataset.

This paper is organized as follows. First, the extension of the existing models with biases is addressed in Section 2. Thereafter, simulation results are presented in Section 3. Section 4 concerns the verification of logical properties against the simulation results obtained to show that the models indeed exhibit the desired behavior. The validation of the model based upon experimental data is presented in Section 5, and finally, Section 6 is a discussion.

II. BIASED HUMAN TRUST MODELS

In order to be able to model human trust, an existing trust model aimed at representing human trust is taken as a basis. This is a well-known model presented in ([9]; see also [13; 14; 15]). The model is expressed as follows:

$$T(t + \Delta t) = T(t) + \gamma\cdot(E(t) - T(t))\cdot\Delta t \qquad \text{(i)}$$

In the trust model, it is assumed that the human receives a certain experience at each time point, *E(t)* and wants to derive a new trust value for the next time point $(t + \Delta t)$. The experience is assumed to reside on the interval [0, 1]. It is then compared with the current trust level (*T(t)*) and the difference is multiplied with a speed factor γ. This difference is then added to the current trust level and results in a new trust level.

The model described above does not include biases, therefore several extensions of the model are introduced in this paper. It is assumed that human biases can affect trust in various ways. First of all, there are different ways in which the bias plays a role in the formation of a new trust value (referred to as the *cognitive dimension*). In this paper, three options are distinguished: (1) the bias solely plays a role in the way in which the human perceives an experience with the specific trustee. In other words, the experience is transformed from a certain objective value to a biased experience (based upon the bias), which is then used to derive a new trust value; (2) the experience is again perceived differently based upon the bias, but the current trust value also plays a role in the perception of the experience; (3) the experiences are not biased, but the trust value itself is biased. Besides these different possibilities of modeling the point at which the bias plays a role in the trust formation, the precise way in which the bias is incorporated within the model can also be varied. There can be a more linear trend in the bias behavior, or it can be of a logistic type. Given these dimensions, in total 6 models for incorporating trust in the unbiased model expressed in equation (i) can now be formulated: 1) linear model with biased experience, 2) linear model with biased experience influenced by current trust, 3) linear model with bias solely determined by current trust, 4) logistic model with biased experience, 5) logistic model with biased experience influenced by current trust, 6) logistic model with bias solely determined by current trust. The above models are abbreviated as LiE, LiET, LiT, LoE, LoET, and LoT respectively.

In order to incorporate the biased behavior in the model presented in equation (i), functions have been defined that take the current experience (for models LiE and LoE), the experience and the trust (for models LiET and LoET), or the trust value itself (models LiT and LoT) and transforms that into a biased value. This biased value can then be used to calculate the new trust value based upon equation (i). For the models that express the bias solely based upon the experience, the following two equations are used (for linear and logistic respectively):

$$\begin{aligned} \textit{LiE: } f(E(t)) &= E(t) + (2\cdot\beta - 1)\cdot(1 - E(t)) && \text{when } \beta > 0.5 \\ &= 2\cdot\beta\cdot E(t) && \text{when } \beta \leq 0.5 \end{aligned}$$

$$\textit{LoE: } f(E(t)) = 1 / [1 + \exp(-\sigma\cdot(E(t) - \tau))]$$

In the first equation, β is the bias parameter for linear transformation which is from interval [0, 1]. Here values for β of 0, 0.5 and 1.0 represent absolute negative, neutral, and absolute positive bias respectively. It can be seen that for the case of a positive bias (i.e. $\beta > 0.5$) the current experience is increased with a factor dependent on the positiveness of the bias (the more positive the bias, the more the objective experience is increased). For the logistic equation (LoE), σ and τ are the steepness and threshold parameters for logistic transformation. In the logistic transformation τ is assumed to represent the human's bias. It is assumed that this value has an inverse relationship with β i.e. $\tau = 1 - \beta$. E(t) and T(t) are the experience and human trust on trustee at time point t respectively. The resulting value of the function f(E(t)) is the biased experience. This can then be incorporated into the base model (equation (i)) as follows:

$$T(t + \Delta t) = T(t) + \gamma\cdot(f(E(t)) - T(t))\cdot\Delta t \qquad \text{(ii)}$$

In the second set of bias equations, the bias plays a role in combination with the current trust value and the experience, as expressed below.

$$\textit{LiET: } f(E(t), T(t)) = \beta\cdot[1 - (1 - E(t))\cdot(1 - T(t))] + (1 - \beta)\cdot[E(t)\cdot T(t)] - T(t)$$

$$\textit{LoET: } f(E(t), T(t)) = 1 / [1 + \exp(-\sigma\cdot(E(t) + T(t) - \tau))] - T(t)$$

The first equation (linear model) expresses that the more positive the bias is, the more the evaluation will be increased depending on the distance of the experience and the trust to the highest value. The second is the logistic variant of the model, whereby the combination of the experience and the trust are used in the threshold function. The function can be inserted into the base model as follows:

$$T(t + \Delta t) = T(t) + \gamma\cdot(f(E(t), T(t)))\cdot\Delta t \qquad \text{(iii)}$$

The final set of equations concerns the bias solely based upon the trust level, and not on the experience itself. The following two equations are used for this purpose:

$$\begin{aligned} \textit{LiT: } f(T(t)) &= T(t) + (1 - T(t))\cdot(T(t) - T(t) + (2\cdot\beta - 1)\cdot(1 - T(t))) && \text{when } \beta > 0.5 \\ &= T(t) + (1 - T(t))\cdot(T(t) - 2\cdot\beta\cdot T(t)) && \text{when } \beta \leq 0.5 \end{aligned}$$

$$\textit{LoT: } f(T(t)) = T(t) + (1 - T(t))\cdot(T(t) - 1 / (1 + \exp(-\sigma\cdot(T(t) - \tau))))$$

The equations follow the same structure as seen for the experience based bias, except that now the trust value is used. It is combined with the base model as follows:

$$T(t + \Delta t) = T(t) + \gamma\cdot(E(t) - f(T(t))\cdot\Delta t \qquad \text{(iv)}$$

Combining the two sets of equations is straightforward and the joint equations are not described in the paper for the sake of brevity.

III. Simulation Results

In order to observe the behavior of bias based trust models described in the previous section, several simulation experiments are performed. The behavior of the models itself is described in Section 3.1. Section 3.2 shows how the models can be used to describe each others behavior.

A. *Single model comparisons*

In this case, merely one agent for which an agent has to form trust is considered. In this section the results of one of these experiments is presented in detail. In Table 1 the experimental configuration for this simulation is described. Here it can be seen that bias parameter is changed from 0.00 to, 0.50 and 1.00 which represents negative, neutral and positive bias respectively. For comparison purposes, the bias parameter τ for the logistic model is calculated by means of the following equation: $\tau = 1 - \beta$. The speed factor γ is taken as 0.25. Furthermore, the initial trust value is taken as 0.5 which means that the human has neutral trust at time point 0. The step size (Δt) is set to0.5. The experiences injected periodically change between the values 0.00, 0.50 and 1.00 respectively with a period of 10 time steps. Each of these experience values represent negative, neutral and positive experience respectively. This experience sequence is used to see the behavior of these models on and between varying extremes. The simulations have been performed using a dedicated program that has been written in C.

Table I. Experimental configuration for Simulation Experiments

Quantity	Symbol	Value
Bias parameter	β (for linear model), τ (logistic model)	0.00, 0.50, 1.00
Rate of change of trust	γ	0.25
Time step	Δt	0.50
Initial trust	T(0)	0.50
Steepness	σ	5
Experiences	E (t)	Periodic (0.0, 0.5, 1.0) on 10 time steps each

In Figure 1-3 the results of the simulations given the experience sequence introduced above are shown. In Figure 1 the agent has a negative bias towards the trustee. A simulation for a neutral bias is shown in Figure 2, whereas a positive bias is used in Figure 3. It can be observed in the case of the negative bias that both the LiE and LiET converge to no trust (value 0) despite the fact that the trustee gives some positive experiences. The LiT, LoT, and LoE variants show almost similar trends compared to the base trust model but with a much lower trust value (which is precisely as desired due to the negative bias). The final variant of the model (LoET) shows an undesired result: the trust is actually higher than the base model. This is due to higher parameter value of parameter σ (steepness) which is 5. For lower values of the steepness (< 3) this model shows desired results as well (but has not been shown for the sake of brevity).

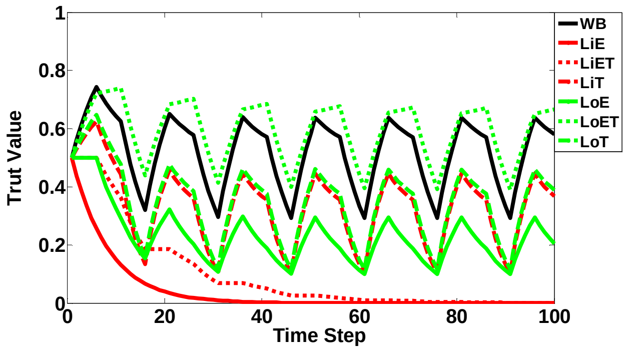


Figure 1. Simulation results for absolute negative bias i.e. (β=0 and τ=1, σ=5)

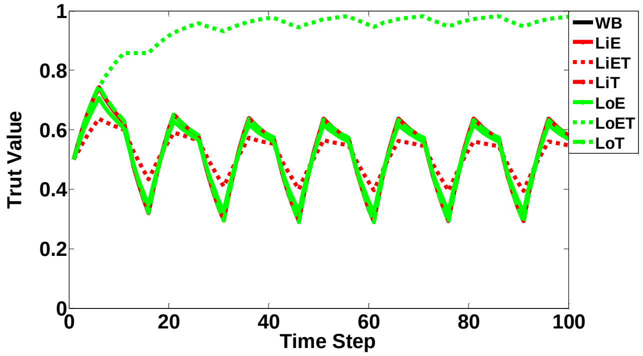


Figure 2. Simulation results for neutral or no bias i.e. (β=0.5 and τ=0.5, σ=5)

In Figure 2 a neutral bias i.e. (β=0.5 and τ=0.5, σ=5) is used, and all the models except for one show behavior similar to the baseline model (which is as expected as there is no bias). The LoET shows very different and undesirable behavior as it converges to maximum trust value.

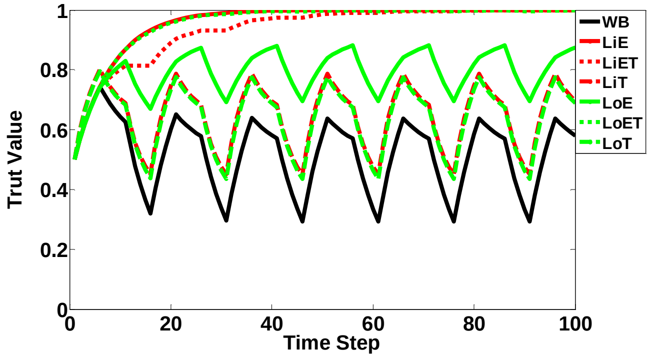


Figure 3. Simulation results for absolute positive bias i.e. (β=1 and τ=0, σ=5)

In Figure 3 an absolute positive bias is set (i.e. β=1 and τ=0, σ=5). In the Figure, the LiE. LiET, and LoET converge to maximum trust (value 1) despite the fact that the trustee gives some negative experiences. This behavior is not completely as desired, but could be adjusted by taking a different steepness value. LoE, LiT and LoT show an almost similar trend as the baseline trust model does, but with higher in trust value, precisely is as desired.

B. *Mutual Mirroring of models*

To analyze the generalization capacity of these models a novel technique named mutual mirroring of models is used [8]. In this method, a specific trace (simulation run) of a source model is taken as a basis, and an exhaustive search within the parameter space of a target model is performed to see how closely the target model can describe the trace of the source model (i.e. what the set of parameters is with minimum error). This gives a good indication how much the models can describe each others behavior, and some indication of similarity. The mirroring is also done in the opposite direction (i.e. from a trace of the target model to parameters of the source model). This process of mirroring both models into each other is called mutual mirroring of models. The mirroring process can provide a good indication on the generalization of models. For more detail, see [8].

TABLE 2. RESULTS FOR MUTUAL MIRRORING OF THE BIAS BASED TRUST MODELS

	Target Model						
S. Model	**LiE**	**LiET**	**LiT**	**LoE**	**LoET**	**LoT**	**AVG**
LiE	0.00	0.04	0.22	0.12	0.14	0.22	**0.12**
LiET	0.02	0.00	0.19	0.10	0.13	0.19	**0.11**
LiT	0.01	0.03	0.00	0.01	0.06	0.00	**0.02**
LoE	0.01	0.03	0.09	0.00	0.08	0.09	**0.05**
LoET	0.03	0.05	0.23	0.11	0.00	0.22	**0.11**
LoT	0.01	0.02	0.00	0.01	0.05	0.00	**0.02**
AVG	**0.02**	**0.03**	**0.12**	**0.06**	**0.08**	**0.12**	

The mirroring techniques have been applied to the models introduced in Section 2. The results are shown in Table 2. Here, the columns represent the target models while the rows represent the source models. For a specific trace of the source model (given a certain set of parameter settings) the parameters of the target model are exhaustively searched to generate behavior similar to the trace of the source model with minimum root mean squared error. The values in each cell of the table represent the average error for nine different source model traces generated with different bias values and experience sequences. In the first row of the table it can be seen that on average the source model LiE can be approximated using the LiE, LiET, LiT, LoE, LoET and LoT variants with error of 0.00, 0.04, 0.22, 0.12, 0.14 and 0.22 respectively. Furthermore in the last column of the first row it can be seen that the average error of the mirroring process with all other models is 0.12. This seems to be the most difficult behavior to approximate on average as the other rows show a lower average value. Especially the behavior of the LiT and LoE can be very well approximated by the other models. Furthermore, in the last row the values are shown that indicate how well a model can describe the other model's behavior. This shows that LiE and LiET can describe many of the source models very well.

IV. VERIFICATION

In this Section, certain desired properties are identified with respect to biased human trust. These properties are verified upon the simulation traces that have been produced by the models proposed in Section 3 to show that the models indeed exhibit this desired behavior. In order to perform this verification in an automated fashion, the language called TTL (for Temporal Trace Language, cf. [2]) has been used. TTL features an automated verification tool that verifies the properties against traces that have been loaded in the verification tool. First, the language TTL is explained in more detail, followed by a section explaining the properties that have been identified for biased trust. Finally, the results of the checks are shown.

A. *Temporal Trace Language (TTL)*

The predicate logical temporal language TTL 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 γ over state ontology Ont, the state in γ at time point t is denoted by state(γ, t). These states can be related to state properties via the formally defined satisfaction relation denoted by the infix predicate |=, i.e., state(γ, t) |= p denotes that state property p holds in trace γ at time t. Based on these statements, dynamic properties can be formulated in a formal manner in a sorted first-order predicate logic, using quantifiers over time and traces and the usual first-order logical connectives such as $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\forall$, $\exists$. For more details on TTL, see [2].

B. *Properties for Biased Trust*

Four properties have been identified with respect to biased behavior of human trust. The first property expresses the general principle of this bias, namely that once a person has a more positive bias towards an agent, this agent will more frequently be the most trusted trustee, as expressed in property P1 below.

P1: General bias property

If within two traces with the same experience sequence in one trace an agent has a more positive bias towards a trustee compared to the other trace, and the agent has the same biases for *the other trustees, then the trustee will more frequently be the trustee with the highest trust value in the trace with the higher bias compared to the trace with the lower bias. This then results in this trustee being requested more frequently.*

P1 ≡

$$\begin{array}{l}
\forall\gamma_1, \gamma_2\text{:TRACE, tr}_1\text{:TRUSTEE, b}_1\text{, b}_2\text{:REAL} \\
[\ \text{same\_experience\_sequence}(\gamma_1, \gamma_2)\ \& \\
\quad \text{state}(\gamma_1, 0) \mid= \text{bias\_for\_trustee}(\text{tr}_1, \text{b}_1)\ \& \\
\quad \text{state}(\gamma_2, 0) \mid= \text{bias\_for\_trustee}(\text{tr}_1, \text{b}_2)\ \&\ \text{b}_1 > \text{b}_2\ \& \\
\quad \forall\text{tr}_2\text{:TRUSTEE} \neq \text{tr}_1\ \exists\text{b}_3\text{:REAL} \\
\qquad [\text{state}(\gamma_1, 0) \mid= \text{bias\_for\_trustee}(\text{tr}_2, \text{b}_3)\ \& \\
\qquad\ \text{state}(\gamma_2, 0) \mid= \text{bias\_for\_trustee}(\text{tr}_2, \text{b}_3)\ ] \Rightarrow \\
[\ \sum_{\forall t\text{:TIME}} \text{case}(\text{highest\_trust\_value}(\gamma_1, \text{t}, \text{tr}_1), 1, 0) \geq \\
\quad \sum_{\forall t\text{:TIME}} \text{case}(\text{highest\_trust\_value}(\gamma_2, \text{t}, \text{tr}_1), 1, 0)\ ]\ ]
\end{array}$$

Where:

same_experience_sequence(γ_1:TRACE, γ_2:TRACE,) ≡

$$\begin{array}{l}
\forall\text{t:TIME, tr:TRUSTEE, v:REAL} \\
[\ \text{state}(\gamma_1, \text{t}) \mid= \text{objective\_experience\_value}(\text{tr}, \text{v}) \Rightarrow \\
\quad \text{state}(\gamma_2, \text{t}) \mid= \text{objective\_experience\_value}(\text{tr}, \text{v})\ ]
\end{array}$$

highest_trust_value(γ:TRACE, t:TIME, tr$_1$:TRUSTEE) ≡

$$\begin{array}{l}
\forall\text{v}_1\text{:REAL} \\
[\ \text{state}(\gamma, \text{t}) \mid= \text{trust\_value}(\text{tr}_1, \text{v}_1) \Rightarrow \\
\quad \forall\text{tr}_2\text{:TRUSTEE} \neq \text{tr}_1\text{, v2:REAL [} \\
\qquad \text{state}(\gamma, \text{t}) \mid= \text{trust\_value}(\text{tr}_2, \text{v}_2) \Rightarrow \text{v}_2 < \text{v}_1\ ]\ ]
\end{array}$$

The second property expresses that the trust level itself will be higher in the case of a more positive bias.

P2: Trust comparison

Trustees for which an agent with a more positive bias have a higher trust value compared to a trace in which the agent has a lower bias with respect to the trustee (given that the experiences are equal as well as the biases for the other trustees).

P2 ≡

$$\begin{array}{l}
\forall\gamma_1, \gamma_2\text{:TRACE, tr:TRUSTEE, b}_1\text{, b}_2\text{:REAL} \\
[\ [\ \text{same\_experience\_sequence}(\gamma_1, \gamma_2)\ \& \\
\quad \text{state}(\gamma_1, 0) \mid= \text{bias\_for\_trustee}(\text{tr}, \text{b}_1)\ \& \\
\quad \text{state}(\gamma_2, 0) \mid= \text{bias\_for\_trustee}(\text{tr}, \text{b}_2)\ \&\ \text{b}_1 > \text{b}_2\ \& \\
\quad \forall\text{tr}_2\text{:TRUSTEE} \neq \text{tr}_1\ \exists\text{b}_3\text{:REAL [} \\
\qquad \text{state}(\gamma_1, 0) \mid= \text{bias\_for\_trustee}(\text{tr}_2, \text{b}_3)\ \& \\
\qquad \text{state}(\gamma_2, 0) \mid= \text{bias\_for\_trustee}(\text{tr}_2, \text{b}_3)\ ] \\
\quad \Rightarrow \\
\quad \forall\text{t:TIME, tv}_1\text{, tv}_2\text{:REAL} \\
\quad [\ \text{state}(\gamma_1, \text{t}) \mid= \text{trust\_value}(\text{tr}, \text{tv}_1)\ \& \\
\quad\ \ \text{state}(\gamma_2, \text{t}) \mid= \text{trust\_value}(\text{tr}, \text{tv}_2)\ ] \Rightarrow \text{tv}_1 \geq \text{tv}_2\ ]
\end{array}$$

In order to facilitate the addition of a bias to existing models, some models transform the experience (i.e. experiences colored by the bias). In case of a more positive bias, the biased experiences will generally be higher (notice that the formalizations have been omitted due to the limited space available).

P3: Experience comparison

The objective experience provided by a trustee is translated into a higher subjective experience for trustees with a higher bias (given the same experience sequence).

Finally, in some of the bias models, trust is explicitly considered to color the experiences. In case the trust level is higher, the same objective experience gets an even more positive value.

P4: Influence of trust upon experience

If the trust level for a certain trustee at time point t is higher than the trust level at another time point t', whereas the objective experience is equal and not on the boundary of the scale (i.e. 0 or 1), then the subjective experience will be higher at time point t.

Note that in the property, the objective experiences on the boundaries are not considered as the influence of trust cannot always be distinguished there (e.g. if an experience of 1 is encountered, the experience can never become higher than 1).

C. Verification Results

The results of the verification are shown in Table 3.It can be seen that property P1 is satisfied for all bias models presented in this paper. When looking at the properties P2 and P3 however, the properties also hold for the various models that have been identified. Finally, property P4 is only satisfied for the models where trust is considered when forming the subjective experience, which makes sense as this property precisely describes this influence. Properties P3 and P4 are actually not relevant for models LoET and LoT as they do not incorporate the notion of *subjective experience*, therefore the property is always satisfied (due to the fact that the antecedent of the implication never holds).

TABLE 3. RESULT OF VERIFICATION

	LiE	LiET	LiT	LoE	LoET	LoT
P1	satisfied	satisfied	satisfied	satisfied	satisfied	satisfied
P2	satisfied	satisfied	satisfied	satisfied	satisfied	satisfied
P3	satisfied	satisfied	satisfied	satisfied	satisfied	satisfied
P4	fails	satisfied	fails	satisfied	satisfied	satisfied

V. VALIDATION

Besides the fact that the models show the desired behavior, the most interesting aspect to see is whether the models describe human behavior better. Therefore, data from a validation experiment (as presented in [5]) has been used to perform a validation, and the results are presented in this section. First, the experiment setup is addressed, followed by the results.

A. Experimental Setup

The experimental task was a classification task in which two participants on two separate personal computers had to classify geographical areas at the same time. These areas had to be classified as areas that either needed to be attacked, helped or left alone by ground troops according to specific criteria. The participants needed to base their classification on real-time computer generated video images that resembled video footage of real unmanned aerial vehicles (UAVs). On the camera images, multiple objects were shown. There were four kinds of objects: civilians, rebels, tanks and cars. The identification of the number of each of these object types was needed to perform the classification. Each object type had a score (either −2, −1, 0, 1 or 2, respectively) and the total score within an area had been determined. Based on this total score the participants could classify a geographical area (i.e., attack when above 2, help when below −2 or do nothing when in between). Participants had to classify two areas at the same time and in total 98 areas had to be classified. Both participants did the same areas with the same UAV video footage. During the time a UAV flew over an area, three phases occurred: The first phase was the advice phase. In this phase both participants and a supporting software agent gave an advice about the proper classification (attack, help, or do nothing). This means that there were three advices at the end of this phase. It was also possible for the participants to refrain from giving an advice, but this hardly occurred. The second phase was the reliance phase. In this phase the advices of both the participants and that of the supporting software agent were communicated to each participant. Based on these advices the participants had to indicate which advice, and therefore which of the three trustees (self, other or software agent), they trusted the most. Participants were instructed to maximize the number of correct classifications at both phases (i.e., advice and reliance phase). The third phase was the feedback phase, in which the correct answer was given to both participants. Based on this feedback (i.e. the experience in the model explained in Section 2) the participants could update their internal trust models for each trustee (self, other, software agent). In Figure 4 the interface of the task is shown.

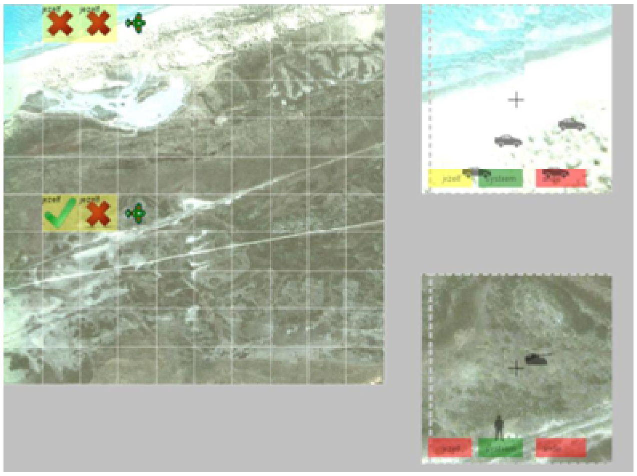


Figure 4. Interface of the experimental task.

B. *Model Evaluation*

In order to compare the different models described in Section 2, the measurements of experienced performance feedback were used as input for the models (i.e. as experiences) and the output (predicted reliance decisions) of the models was compared with the actual reliance decisions of the participant. It is hereby assumed that the human always consults the most trusted trustee. Hence, the reliance decision indicates which trustee is trusted most. Of course, the model still has a number of parameters that need to be tuned towards the specific participant. Therefore, an exhaustive search approach has been taken to tune the parameters of the trust model (cf. [6]). The resulting set of parameters is the set with minimum error in the prediction of the reliance decisions for that specific participant. Hence, the relative overlap of the predicted and the actual reliance decisions was a measure for the accuracy of the models. As these models have a different number of parameters the parameter tuning process took a different amount of time for each of different models. Assuming that S is the number of subjects, M number of model types (namely unbiased, linear and logistic), B number of bias types (using experience, trust, and experience and trust), P the number of parameters with α degree of precision of the parameters (in the range of 0 - 1), T the number of time steps, and N number of trustees, the complexity is then O $(S.M.B.10^{P\alpha}.T.N)$. This indicates that it is exponential in the number of parameters and their precision value. Models presented here have different number of parameter with different types of precisions. The baseline model has one parameter γ (with 0.01 precision), which is assumed to be the same for all trustees. The linear models have four (γ, β_1, β_2, β_3 with 0.01) where β_1, β_2, and β_3 represent the bias of the subject towards each trustee and the logistic models have seven parameter (γ, τ_1, τ_2, τ_3, σ_1, σ_2, and σ_3, where γ and τ has precision 0.01 and σ has precision 1 within range 1 to 20). In order to enable the parameter estimation to be done within a reasonable time, the DAS-4 cluster has been used [1]. It took a total of 6.19 hours to run on the DAS cluster, whereas on a result computer it would have cost 166.66 days (based on the complexity function).

The results of the validation process are in the form of accuracies per trust model (unbiased model (UM), LiE, LiT, LiET, LoE, LoT, LoET and the best fit model (MAX)). The differences in accuracy are detected using a repeated measures analysis of variance (ANOVA) and post-hoc Bonferroni t-tests. Following to this, to test for robustness, the best fit model is also cross-validated (i.e., tuning on half the data (known data), and validating on the other half (unknown data), and vice versa) against the unbiased model, using a paired t-test.

C. *Validation Results*

From the data of 18 participants (eight male and ten female, with an average age of 23 (SD = 3.8)), two outliers have been removed, which leaves a data set of 16 accuracies per model type (UM, LiE, LiT, LiET, LoE, LoT, LoET and MAX). In Fig. 5a the subjects are shown on the x-axis while the prediction accuracies of the models are presented on y-axis. Here it can be seen that the LiE and LoET variants are mostly on the upper bound of the prediction accuracy whereas the LiT, LiET, and LoT are on the lower bound. In Fig. 5b the average accuracy of the models over the participant is shown. It can be seen that the LiE and LoET variant provide better predictions while the LiT, LiET, LoE, and LoT perform worse compared to the baseline model (UM).

In Figure 6 the main effect of model type for accuracy for known data is shown. A repeated measures analysis of variance (ANOVA) showed a significant main effect (F(7, 105) = 61.04, p << .01). A post-hoc Bonferroni test showed that there is a significant difference between all biased model types and the unbiased model (UM), p << 0.01, for all tests. For models UM, LiT, LiET and LoT a significantly higher accuracy was found for the best fit model (MAX), p << 0.01, for all tests.

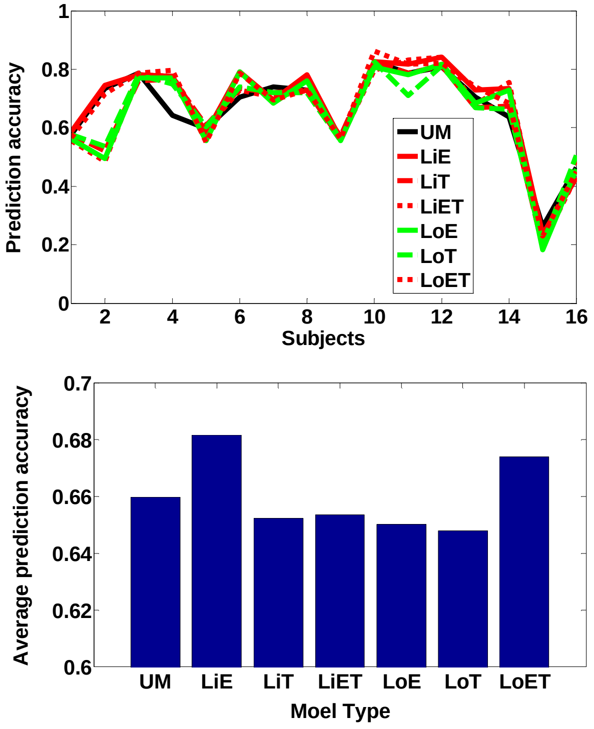
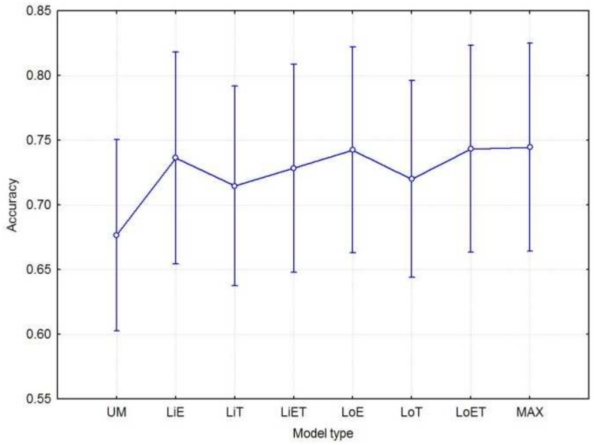


Figure 5. a) prediction accuracy of models across subjects,



b) average prediction accuracy of models for all subjects

Figure 6. Main effect of model type on accuracy.

Finally, for unknown data, a paired t-test showed a significant improved accuracy of the best fit model (M=0.70, SD=0.16) compared to the unbiased model (M=0.66, SD=0.15), t(15)=3.13, p<<0.01. This means that at least one of the different biased models shows an increased capability to estimate trust of the tested participants compared to the model without an explicit bias incorporated, also for unknown data.

VI. Discussion

In this paper, an approach has been presented that allows for the modeling biases in human trust behavior. In order to come to such an approach, an existing model (cf. [9]) has been extended with additional concepts. A number of different variants have hereby been introduced: (1) a model that strictly places the bias on the experience obtained from the trustee (2) a model that combines the trust and experience and then applies the bias, and (3) a model that uses the previous trust value on which the bias is applied. Simulation results of the behavior of each of the model have been shown, as well as a comparison of the behavior of the models via mutual mirroring (cf. [8]). Furthermore, the resulting patterns have been verified by means of the verification of formal desired properties and were shown to behave as expected.

Of course, the main goal of the research is to model human trust. Therefore, an extensive validation has taken place in which the trust models are used to describe and forecast human trust levels. Hereby, parameter estimation techniques are utilized to tailor the model towards the behavior of the participant. In this paper, a simple parameter estimation technique has been used, but more complex estimation techniques could also be applied. The validation study showed that for each participant at least one of the different biased models has an increased capability to estimate trust, also for unknown data. For known data (i.e., the models were tuned to it), all of the models are better compared to the tuned unbiased model. The latter means that if one is able to develop a kind of on-line tuning, the accuracies of the models would certainly benefit. The first means that the identification of personal characteristics might lead to an online form of the selection of the best fit model for unknown data, which on its turn leads to an improved accuracy.

More models that represent human trust exist in the literature (see e.g. [10; 11; 4]). Given the approach presented in this paper, these models could also be extended with biases. It is part of future work to see whether these extensions would also improve the accuracy of those models.

Acknowledgment

This research was partly funded by the Dutch Ministry of Defense under program number V929. Furthermore, this research has partly been conducted as part of the FP7 ICT Future Enabling Technologies program of the European Commission under grant agreement no. 231288 (SOCIONICAL).

REFERENCES

[1] Bal, H., Bhoedjang, R., Hofman, R., Jacobs, C., Kielmann, T., Maassen, J., et al. (2000). The distributed ASCI Supercomputer project. SIGOPS Oper. Syst. Rev., 34(4), 76-96.

[2] Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A., and Treur, J., Specification and Verification of Dynamics in Agent Models. *International Journal of Cooperative Information Systems*, vol. 18, 2009, pp. 167 - 193.

[3] Danek, A., Urbano, J., Rocha, A.P., and Oliveira, E. (2010). Engaging the Dynamics of Trust in Computational Trust and Reputation Systems. In: P. Jedrzejowicz et al. (eds.): Proc. KES-AMSTA 2010, Part I, LNAI 6070, 2010, pp. 22–31.

[4] Falcone R., and Castelfranchi, C. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In: Proceedings of AAMAS 2004, pp. 740–747, 2004.

[5] Hoogendoorn, M., Jaffry, S.W., and van Maanen, P.P., Validation and Verification of Agent Models for Trust: Independent compared to Relative Trust. In: Proceedings of the 5th IFIP WG 11.11 International Conference on Trust Management, to be published by Springer-Verlag.

[6] Hoogendoorn, M., Jaffry, S.W., and Treur, J., An Adaptive Agent Model Estimating Human Trust in Information Sources. In: Baeza-Yates, R., Lang, J., Mitra, S., Parsons, S., Pasi, G. (eds.), *Proc. of the 9th IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology, IAT'09*. IEEE C.S. Press, 2009, pp. 458-465.

[7] Huff, L., and Kelley, L., Levels of Organizational Trust in Individualist versus Collectivist Societies: A Seven Nation Study. *Organizational Science*, vol. 14, pp. 81-90.

[8] Jaffry, S.W., Treur, J.: Comparing a Cognitive and a Neural Model for Relative Trust Dynamics. In: Leung, C., Lee, M., Chan, J. (eds.): Neural Information Processing, Vol. 5863. Springer Berlin / Heidelberg (2009) 72-83

[9] Jonker, C.M., and Treur, J., Formal Analysis of Models for the Dynamics of Trust based on Experiences. In: F.J. Garijo, M. Boman (eds.), Multi-Agent System Engineering, In: 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99. Lecture Notes in AI, vol. 1647, Springer Verlag, 1999, pp. 221-232.

[10] Ramchurn, S.D., Huynh, D., Jennings, N.R., Trust in Multi-Agent Systems, The Knowledge Engineering Review, vol. 19, 2004, pp. 1-25.

[11] Sabater, J., and Sierra, C., Review on Computational Trust and Reputation Models, A.I. Review, vol. 24, 2005, pp. 33-60.

[12] Sears, D.O., The Person Positivity Bias. *Journal of Personality and Social Psychology*, vol. 44, 1983, pp. 233-250.

[13] Singh, S.I., and Sinha, S.K. (2010). A New Trust Model Based on Time Series Prediction and Markov Model. In: Das, V.V., and Vijaykumar, R. (eds.), Proc. of the Information and Communication Technologies International Conf., ICT 2010, CCIS, volume 101, Springer Verlag, 2010, pp. 148-156.

[14] Skopik, F., Schall, D., Dustdar, S. (2010). Modeling and mining of dynamic trust in complex service-oriented systems, Information Systems 35 (2010) 735–757.

[15] Walter, F.E., Battiston, S., and Schweitzer. F., (2009). Personalised and Dynamic Trust in Social Networks. In: Bergman, L, Tuzhilin, A., Burke, R., Felfernig, A., Schmidt-Thieme, L., Proceedings of the Third ACM conference on Recommender systems, RecSys’09, ACM, pp. 197-204.

[16] Yamagishi, T., Jin, N, and Miller, A.S. In-group bias and culture of collectivism. *Asian Journal of Social Psychology*, vol. 1, pp. 315-328.