Modelling Biased Human Trust Dynamics[1]

Mark Hoogendoorn[*], Syed Waqar Jaffry[*,x], Peter-Paul van Maanen[*,+], and Jan Treur[*]
[*]*VU University Amsterdam, Agent Systems Research Group, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands. {mhoogen, swjaffry, treur}@few.vu.nl*
[x]*Punjab University College of Information Technology (PUCIT), University of The Punjab, Shahrah-e-Quaid-i-Azam, Lahore, Pakistan. swjaffry@pucit.edu.pk*
[+]*Netherlands Organisation for Applied Scientific Research (TNO), Department of Perceptual and Cognitive Systems, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands peter-paul.vanmaanen@tno.nl*

Abstract. Within human trust related behaviour, according to the literature from the domains of Psychology and Social Sciences often non-rational behaviour can be observed. Current trust models that have been developed typically do not incorporate non-rational elements in the trust formation dynamics. In order to enable agents that interact with humans to have a good estimation of human trust, and take this into account in their behaviour, trust models that incorporate such human aspects are a necessity. A specific non-rational element in humans is that they are often biased in their behaviour. In this paper, models for human trust dynamics are presented incorporating human biases. In order to show that they more accurately describe human behaviour, they have been evaluated against empirical data, which shows that the models perform significantly better.

Keywords. Trust, Biases, Modelling, Validation.

[1] The work presented in this paper is a significant extension by more than 40% of (Hoogendoorn, Jaffry, Maanen, and Treur, 2011).

1. Introduction

Within the domain of multi-agent systems, a variety of trust models have been proposed (e.g., see [13], [14] for an overview). Often, such trust models are utilized in an environment in which software agents should make choices based upon their levels of trust, and hence, such models aim to optimize the behavior of the agent by using the most appropriate trust function. An example of such a model is for instance described in [12]. In situations where software agents interact with humans, trust models that are incorporated in these agents may have a completely different purpose: to estimate the trust levels of the human over time, and take that into consideration in its behavior, for example, by providing advices from other trustees that are trusted more. If this is the purpose of the trust model, then the model should also explicitly incorporate non-rational human aspects. Examples of models taking into account various human aspects are [3], [7], [11].

In the literature in the domain of Psychology and Social Sciences it has been shown that one important non-rational aspect within the formation of trust is the incorporation of biases. Several biases have been observed whereby the culture bias is one of the most reported ones. In [20] it is shown that humans from collectivistic cultures tend to have a bias towards trusting members that belong to the same group and avoid the persons from outside the group. In [8] also a comparison between individualistic and collectivistic cultures is made which shows that the trust of the members of an individualistic society is less negatively biased towards persons from outside their group. Other authors also emphasize the existence of such a bias in general, e.g. [15]. 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 this paper, a model has been developed that incorporates biases in a model for trust dynamics. In order to do so, an existing trust model is taken as a point of departure (cf. [11]), which was applied, for example, in [17], [18], [19]). Biases have been added to this model using a number of different approaches for the manner in which biases affect the level of trust. Introducing a trust model with the purpose to model human behaviour in a more realistic way requires a thorough evaluation of the model. Therefore in this paper, a number of approaches have been used to evaluate the introduced models. First of all, the behaviour of the models themselves have been rigorously compared and analyzed using identified emerging properties. Also, an extensive mathematical analysis of monotonicity, equilibria and behaviour around equilibria has been performed for this purpose. In addition to these types of formal analysis, also an empirical analysis has been performed. The models have been validated against empirical data that has been obtained from an experiment conducted with human subjects. Such a full empirical validation is not so common for computational trust models. However, some authors have done some form of validation. For instance, in (Jonker, Schalken, Theeuwes and Treur 2004) an experiment has been conducted whereby the trends in human trust behaviour have been analyzed to verify properties underlying trust models developed in the domain of multi-agent systems. However, no attempt was made to exactly fit the model to the trusting behaviour of the human. The outcome of the validation experiment presented in the current paper shows that the introduced bias-based models perform significantly better than comparable models without explicit representation of biases.

This paper is organized as follows. First, in Section 2 six new human bias-based trust models are introduced across computational and human cognitive dimensions. Thereafter, simulation results of these bias-based trust models are presented in Section 3. The formal analyses of the newly designed bias-based trust models through logical and mathematical means are described in Section 4 and 5, respectively. Thereafter, the human-based trust experiment is explained in Section 6. The validation results of the models based on empirical data collected in the experiment described in Section 6 are presented in Section 7, and finally, Section 8 is a discussion.

2. Models for Biased Trust Dynamics

In this section a number of trust models are proposed that incorporate biased human behaviour. In order to be able to model bias-based trust dynamics, an existing trust model aimed at representing human trust is taken as a basis. This is a well-known model presented in [11] and applied, for example, in [17], [18], [19]. The model is expressed as follows:

$$T(t + \Delta t) = T(t) + \gamma \left(E(t) - T(t)\right) \Delta t \qquad (1)$$

In this trust model, it is assumed that the human receives a certain experience at each time point, *E(t)*. The experience is represented by a value in the interval [0, 1]. It is then compared with the current trust level *T(t)* and the difference is multiplied with a trust update speed factor γ. This difference is then multiplied by the chosen step size Δt and added to the current trust level to obtain a new trust level.

The model described above does not include biases; therefore in this paper extensions of the model are introduced incorporating biases. This can be done in different manners. It is assumed that human biases can affect trust in a number of ways. More specifically, there are different ways in which the bias plays a role in the formation of a new trust value; this is referred to as the *cognitive dimension* in Fig. 1. In this paper, three options are distinguished:

(a) the bias solely plays a role in the way in which the human perceives an experience with a specific trustee: the experience is transformed from a certain objective value to a subjective biased experience value, which is then used to derive a new trust value.
(b) 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.
(c) the experiences are not biased, but the trust value itself is biased.

Besides these different possibilities of modelling the point at which the bias plays a role in the trust formation process, the precise way in which the bias is incorporated within the model can also be varied. There can be assumed a more linear trend in the bias behaviour, or a logistic type of trend can be assumed; this is referred to as the *computational dimension* in Fig. 1. Given these dimensions, in total 6 models for incorporating trust in the unbiased model expressed in equation (1) can now be formulated (see Fig. 1):

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 behaviour in the model presented in equation (1), 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 (for 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 (1).

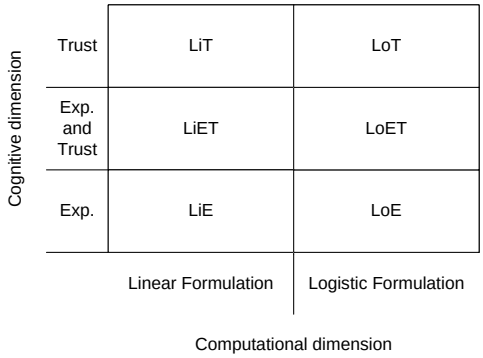


Fig. 1. Bias-based trust models

2.1 Trust models with biased experience

For the models that express the bias solely based upon the experience, the following two equations are used (for linear and logistic respectively):

LiE:
$f\big(E(t)\big) = E(t) + (2\beta - 1)\big(1 - E(t)\big)$ *when* $\beta > 0.5$
$f\big(E(t)\big) = 2\beta E(t)$ *when* $\beta \leq 0.5$
LoE:
$f\big(E(t)\big) = 1/\big(1 + e^{(-\sigma(E(t)-\tau))}\big)$

In the first equation, β is the bias parameter from the interval [*0, 1*]. Here values for β of *0.0*, *0.5* and *1.0* represent an 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 the 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. $\tau = 1 - \beta$). Furthermore *E(t)*, and *T(t)* are the experience and human trust level on the given trustee at time point *t*, respectively. The resulting value of the function *f(E(t))* is the biased experience.

This function can be incorporated into the base model (equation (1)) in a general setting as follows:

$$T(t + \Delta t) = T(t) + \gamma \left(f\big(E(t)\big) - T(t)\right)\Delta t \qquad (2)$$

For the specific (linear and logistic) cases considered this becomes:

$$T(t+\Delta t) = T(t) + \gamma\Big(E(t) + (2\beta - 1)\big(1 - E(t)\big) - T(t)\Big)\Delta t \quad when\ \beta \geq 0.5$$

$$T(t+\Delta t) = T(t) + \gamma\big(2\beta E(t) - T(t)\big)\Delta t \quad when\ \beta \leq 0.5$$

$$T(t+\Delta t) = T(t) + \gamma\Big(1/\big(1 + e^{(-\sigma(E(t)-\tau))}\big) - T(t)\Big)\Delta t$$

2.2 Trust models with biased experience affected by current trust

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.

LiET:

$$f\big(E(t),T(t)\big) = \beta\Big(1 - \big(1 - E(t)\big)\big(1 - T(t)\big)\Big) + \\ +(1-\beta)E(t)T(t) - T(t)$$

LoET:

$$f\big(E(t),T(t)\big) = 1/\big(1 + e^{(-\sigma(E(t)+T(t)-\tau))}\big) - T(t)$$

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 in a general setting as follows:

$$T(t+\Delta t) = T(t) + \gamma\Big(f\big(E(t),T(t)\big)\Big)\Delta t \qquad (3)$$

For the specific (linear and logistic) cases considered this becomes:

$$T(t+\Delta t) = T(t) + \gamma\Big(\beta\Big(1 - \big(1 - E(t)\big)\big(1 - T(t)\big)\Big) + (1-\beta)E(t)T(t) - T(t)\Big)\Delta t$$

$$T(t+\Delta t) = T(t) + \gamma\Big(1/\big(1 + e^{(-\sigma*(E(t)+T(t)-\tau))}\big) - T(t)\Big)\Delta t$$

2.3 Trust models with bias solely determined by current trust

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:

LiT:

$$f\big(T(t)\big) = T(t) + \big(1 - T(t)\big)(1 - 2\beta)\big(1 - T(t)\big) \quad when\ \beta > 0.5$$

$$f\big(T(t)\big) = T(t) + \big(1 - T(t)\big)(1 - 2\beta)T(t) \quad when\ \beta \leq 0.5$$

LoT:

$$f\big(T(t)\big) = T(t) + \big(1 - T(t)\big)\big(T(t) - 1/\big(1 + e^{(-\sigma(T(t)-\tau))}\big)\big)$$

The equations follow the same structure as seen for the experience-based bias, except that now the trust value is used.

For the general setting it is combined with the base model as follows:

$$T(t+\Delta t) = T(t) + \gamma\Big(E(t) - f\big(T(t)\big)\Big)\Delta t \qquad (4)$$

For the specific (linear and logistic) cases considered this becomes:

$$T(t+\Delta t) = T(t) + \gamma\Big(E(t) - \Big(T(t) + \big(1 - T(t)\big)(1 - 2\beta)\big(1 - T(t)\big)\Big)\Big)\Delta t \quad when\ \beta \geq 0.5$$

$$T(t+\Delta t) = T(t) + \gamma\Big(E(t) - \big(T(t) + \big(1 - T(t)\big)(1 - 2\beta)T(t)\big)\Big)\Delta t \quad when\ \beta \leq 0.5$$

$$T(t+\Delta t) = T(t) + \gamma\Big(E(t)\Big(T(t) + \big(1 - T(t)\big)\big(T(t) - 1/\big(1 + e^{(-\sigma(T(t)-\tau))}\big)\big)\Big)\Big)\Delta t$$

3. Simulation Results for the Biased Human Trust Models

In order to observe the behaviour of bias-based trust models described in the previous section, several simulation experiments are performed. In these simulation experiments first each model is simulated independently against a set of experience values and then these models are compared using a novel technique called mutual mirroring of models as described in [9].

3.1 Single model comparisons

In this first experiment, merely one trustee 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 *0.0* to, *0.5* and *1.0* 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 trust rate change γ is taken as *0.25*. Furthermore, the initial trust value is taken as *0.50* which means that the human has neutral trust at time point *0*. The step size (Δt) is set to *0.50*.

Table 1: Experimental configuration for simulation experiments

Quantity	**Symbol**	**Value**
Bias parameter	β (linear model) τ (logistic model)	0.00, 0.50, 1.00 1.00, 0.50, 0.00
Trust change rate	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

The experience sequence used in this experiment is represented in Fig. 2. It can be seen that experience provided in this experiment change periodically between the values *0.0*, *0.5* and *1.0* respectively with a period of *10* time steps. Each of these experience values represents negative, neutral and positive experience respectively. This experience sequence is used to see the behaviour of these models on and between varying extremes.

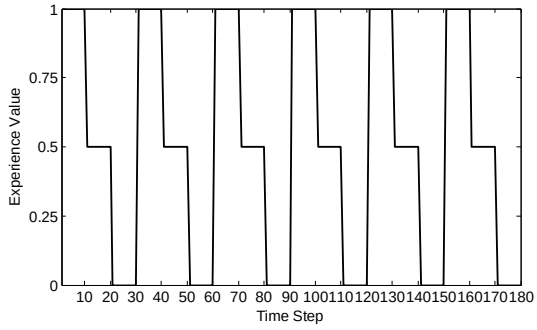


Fig. 2. Experience sequence

In Figures 3-5 the results of the simulations given the experience sequence introduced above are shown.

In Fig. 3 the agent has a negative bias towards the trustee. A simulation for a neutral bias is shown in Figure 4, whereas a positive bias is used in Figure 5. 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).

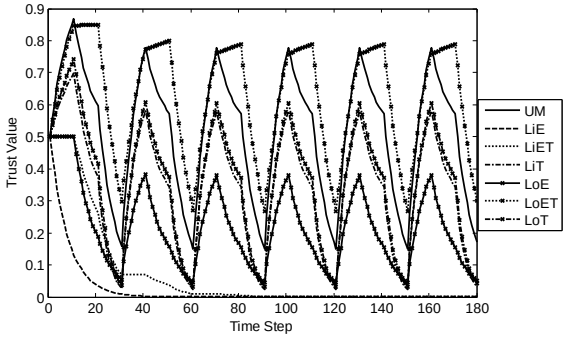


Fig. 3. Simulation results for absolute negative bias ($\beta=0$ and $\tau=1$, $\sigma=5$)

In Fig. 4 a neutral bias i.e. ($\beta=0.5$ and $\tau=0.5$, $\sigma=5$) is used, and all the models except for one show behaviour similar to the baseline model (which is as expected as there is no bias). The LoET does however show very different and undesirable behaviour as it converges to maximum trust value. This relates to the fact that for this type of model the value *0.5* does not show an upward-downward symmetry as required for a non-biased case. Therefore this model does not qualify well in this respect.

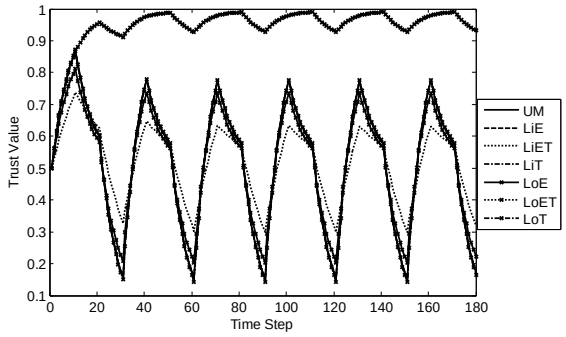


Fig. 4. Simulation results for neutral or no bias ($\beta=0.5$ and $\tau=0.5$, $\sigma=5$)

In Fig. 5 an absolute positive bias is set (i.e. $\beta=1$ and $\tau=0$, $\sigma=5$). 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.

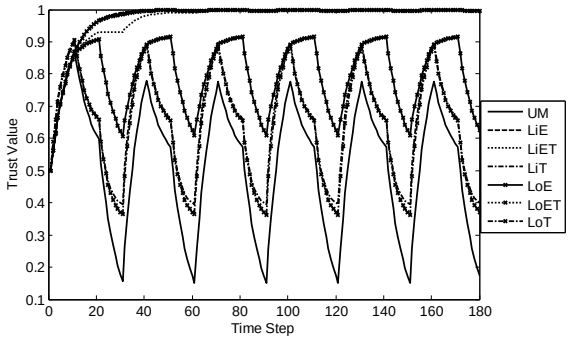


Fig. 5. Simulation results for absolute positive bias ($\beta=1$ and $\tau=0$, $\sigma=5$)

3.2 Mutual mirroring of the bias-based trust models

To analyze the generalization capacity of these models a novel technique named mutual mirroring of models is used as introduced in [9]; see also [7]. In this method, a specific trace (simulation run) of a source model is taken as a basis, and a parameter tuning approach (e.g., exhaustive search within the parameter space) for 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' behaviour, 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 similarity of models. For more details on the approach see [7], [9].

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 behaviour 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 behaviour to approximate on average as the other rows show a lower average value. Especially the behaviour 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 behaviour. This shows that LiE and LiET can describe many of the source models very well.

Table 2. Results for mutual mirroring of the models

		Target Model						
Source 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**	

4. Logical Verification of the Bias-based Trust Models

When developing a new model, a thorough analysis of the behaviour is required to have sufficient confidence in the appropriate behaviour of the model. One way to perform such an analysis is to conduct a mathematical analysis (see Section 5). However, given the complexity of the models proposed in this paper, the analysis of more complex (temporal) patterns might not be feasible using these techniques. Therefore, in this section, certain desired emergent properties are discussed with respect to the bias-based trust models that express complex patterns over time. To show that the models indeed generate this desired behaviour, these properties have been verified upon the simulation traces that have been produced by the models proposed in Section 2. This does not prove a complete adherence of the model to

the properties, but it does shown that for the selected simulation runs (which are of course carefully selected in order to have representative results) adhere to the properties or not In order to perform this verification in an automated fashion, the hybrid temporal language TTL (Temporal Trace Language, cf. [2], [16]) and its software environment has been used. In addition to a dedicated editor TTL features an automated verification tool that automatically verifies specified properties against traces that have been loaded in the verification tool. The language TTL is explained first, followed by a presentation of the desired properties related to trust.

4.1 Temporal Trace Language (TTL)

The hybrid 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 $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\forall$, $\exists$. As a built-in construct in TTL, summations can be expressed, indexed by elements X of a sort S:

$$\Sigma_{X:S}\ \text{case}(\varphi(X), V1, V2)$$

Here for any formula $\varphi(X)$, the expression

$$\text{case}(\varphi(X), V1, V2)$$

indicates the value V1 if $\varphi(X)$ is true, and V2 otherwise. For example,

$$\Sigma_{X:S}\ \text{case}(\varphi(X), 1, 0)$$

simply denotes the number of elements X in S for which $\varphi(X)$ is true. As expressing counting and summation in a logical format in an elementary manner in general leads to rather complex formulae, this built-in construct is very convenient in use. For more details on TTL and the precise functioning of the checker tool, see [2], [16].

4.2 Verification of Bias-based Trust Models

This section describes verification process for the bias-based trust models presented in Section 2. First, in Section 4.2.1 the properties that have been identified for bias-based trust models are introduced and then in Section 4.2.2 results of the checks are presented.

4.2.1 Properties for bias-based trust models

Four properties have been identified with respect to biased behaviour of human trust. The first property expresses the general principle of the bias, namely that once a person has a more positive bias towards a trustee, this trustee will more frequently be the most trusted trustee, as expressed in property P1 below. Note that in this property (and also for properties P2 and P3), it is assumed that the bias does not change during the simulation, and hence, the value at the first time point is selected.

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. For example, this then results in this trustee being selected more frequently.

The formalization of the property is shown below. First, it is checked whether the traces that are being compared contain the same experience sequence. Furthermore, it is checked whether the biases for the trustee tr_1 considered differ (and in fact, is higher in the first trace). Note that this comparison is done at time point 0 as it is assumed that the bias does not change over time in a single run. Furthermore, it is checked whether there exists a single bias value the agents has for all other trustees in both traces, then you sum the cases where the trustee tr_1 is the trustee with the highest trust value and this amount should be higher in the first trace compared to the second.

P1 $\equiv \forall \gamma_1, \gamma_2$:TRACE, tr_1:TRUSTEE, b_1, b_2:REAL
[same_experience_sequence(γ_1, γ_2) &
state(γ_1, 0) |= bias_for_trustee(tr_1, b_1) &
state(γ_2, 0) |= bias_for_trustee(tr_1, b_2) & $b_1 > b_2$ &
$\forall tr_2$:TRUSTEE $\neq tr_1$ $\exists b_3$:REAL
[state(γ_1, 0) |= bias_for_trustee(tr_2, b_3) &
state(γ_2, 0) |= bias_for_trustee(tr_2, b_3)] $\Rightarrow$

$[\sum_{t:TIME}$ case(highest_trust_value(γ_1, t, tr_1), 1, 0) ≥

$\sum_{t:TIME}$ case(highest_trust_value(γ_2, t, tr_1), 1, 0)]]

Here the same experience sequence is simply a property expressing that the experience values in both traces should be the same:

same_experience_sequence(γ_1:TRACE, γ_2:TRACE,) ≡
∀t:TIME, tr:TRUSTEE, v:REAL
[state(γ_1, t) |= objective_experience_value(tr, v) ⇒
state(γ_2, t) |= objective_experience_value(tr, v)]

In the formalisation of the predicate indicating the highest trust value which is used in P1 the trust value for the trustee considered is bound by the ∀-quantifier. For this value it is then checked whether for all other trustees and trust values encountered no higher value than the value for trustee tr_1 is encountered.

highest_trust_value(γ:TRACE, t:TIME, tr_1:TRUSTEE) ≡
∀v_1:REAL
[state(γ, t) |= trust_value(tr_1, v_1) ⇒
∀tr_2:TRUSTEE ≠ tr_1, v2:REAL
[state(γ, t) |= trust_value(tr_2, v_2) ⇒ $v_2 < v_1$]]

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 has 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).

The formalization of this property is very similar to P1, except that now a comparison is made between the trust values themselves.

P2 ≡ ∀γ_1, γ_2:TRACE, tr_1:TRUSTEE, b_1, b_2:REAL
[same_experience_sequence(γ_1, γ_2) &
state(γ_1, 0) |= bias_for_trustee(tr_1, b_1) &
state(γ_2, 0) |= bias_for_trustee(tr_1, b_2) & $b_1 > b_2$ &
∀tr_2:TRUSTEE ≠ tr_1 ∃b_3:REAL
[state(γ_1, 0) |= bias_for_trustee(tr_2, b_3) &
state(γ_2, 0) |= bias_for_trustee(tr_2, b_3)] ⇒
∀t:TIME, tv_1, tv_2:REAL
[state(γ_1, t) |= trust_value(tr_1, tv_1) &
state(γ_2, t) |= trust_value(tr_1, tv_2)] ⇒ $tv_1 ≥ tv_2$]

In order to facilitate the addition of bias to existing models, a translation scheme has been proposed to translate objective experiences into subjective experiences (i.e., experiences coloured by the bias). In case of a more positive bias, the biased experiences will be at least as high.

P3: Experience comparison
The objective experience provided by a trustee is translated into a higher subjective experience for trustees for which the agent has a higher bias (given the same experience sequence).

The formalization of this property takes the first part which is by now well-known from P1 and P2 as n antecedent and checks to see whether the subjective experiences are indeed at least as high for the trace in which a higher bias is encountered.

P3 ≡ ∀γ_1, γ_2:TRACE, tr:TRUSTEE, b_1, b_2:REAL
[[same_experience_sequence(γ_1, γ_2) &
state(γ_1, 0) |= bias_for_trustee(tr, b_1) &
state(γ_2, 0) |= bias_for_trustee(tr, b_2) & $b_1 > b_2$] ⇒
∀t:TIME, ev_1, ev_2:REAL
[[state(γ_1, t) |= subjective_experience_value(tr, ev_1) &
state(γ_2, t) |= subjective_experience_value(tr, ev_2)]
⇒ $ev_1 ≥ ev_2$]]

Finally, in some of the bias model, trust is explicitly considered to colour 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*.

The formalization of this property is a bit more complicated. First, the property binds the trust value at a time point at time point t for a certain trustee as well as the objective experience. Hereby, a check is performed to make sure the objective experience is neither 0 nor 1 as this would sometimes make it impossible to have a higher subjective value. Given that this is the case, and given that the objective experience is the same at another time point t' at which the trust value is lower compared to the trust value at time t, this means that the subjective value at time t must be higher.

P4 ≡ ∀γ:TRACE, t, t':TIME, tr:TRUSTEE,
tv_1, tv_2, ov, sv_1, sv_2:REAL
[state(γ, t) |= trust_value(tr, tv_1) &

state(γ, t) |= objective_experience_value(tr, ov) &
ov > 0 & ov < 1 &
state(γ, t) |= subjective_experience_value(tr, sv_1) &
state(γ, t') |= trust_value(tr, tv_2) & $tv_1 > tv_2$ &
state(γ, t') |= objective_experience_value(tr, ov) &
state(γ, t') |= subjective_experience_value(tr, sv_2)]
$\Rightarrow$ $sv_1 > sv_2$

4.2.2 Verification results for bias-based trust models

Based upon the traces resulting from simulations of the trust models so-called traces have been generated. These traces are essentially logs of the simulations that indicate for each time point what states hold. These traces are loaded into the TTL Checker software which then expresses whether a property (i.e. P1-P4) holds for the trace (or a combination of traces) or not. 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	failed	satisfied	failed	satisfied	satisfied	satisfied

5. Mathematical Analysis of Bias-based Trust Models

The models explored in this paper are adaptive with respect to the experiences of the agent. This means, for example, that when in a time period with very positive experiences, also trust will reach higher levels, and in periods with less positive experiences trust levels will go down. For very long periods of experiences of the same level, the trust level will reach some stable level, which is an equilibrium for the model for the given experience level. It gives a more depended insight in the model when it is known what the value of such an equilibrium is for a given experience level: the model will drive the trust level in the direction of that value. Moreover, the speed by which such a convergence process takes place also is useful information about a model. For these types of analyses the techniques used in the previous section are not practical to use, but mathematical techniques are available that can be used quite well.

The properties addressed here by such mathematical techniques focus for a given point in time *t* in particular on criteria that determine whether due to a given experience the trust level will increase, decrease or will be in equilibrium. Moreover for the equilibria of the models, the behaviour near such equilibria is addressed: whether they are attracting or not, and how fast the convergence takes place. These properties are much more specific and limited compared to the wider types of properties addressed in Section 4, but the mathematical methods allow for more in depth results.

First the general case is addressed; in Table 4 an overview of the results for the general case is summarised. Next, the analysis is made more specific for the case of linear functions; at the end of the section in Table 5 an overview of the results for these specific linear functions is presented. Note that the analysis is done for any given time point *t*, which is sometimes indicated as an argument, but will sometimes be left out to get expressions more transparent.

5.1 Mathematical analysis of trust models with biased experience

Recall that for the models that express the bias solely based upon the experience, the following difference equation is used.

$$T(t + \Delta t) = T(t) + \gamma \left(f\big(E(t)\big) - T(t) \right) \Delta t$$

where it is assumed that $\gamma > 0$. Note that from the equation above it immediately follows:

$$T(t + \Delta t) = T(t) \leftrightarrow f\big(E(t)\big) - T(t) = 0$$
$$T(t + \Delta t) > T(t) \leftrightarrow f\big(E(t)\big) - T(t) > 0$$
$$T(t + \Delta t) < T(t) \leftrightarrow f\big(E(t)\big) - T(t) < 0$$

So, in this case the following criteria can be obtained for trust models with biased experiences:

Equilibrium, increasing and decreasing: trust models with biased experiences

(a) T is in equilibrium for a given E if and only if $f(E) = T$
(b) T is increasing if and only if $f(E) > T$
(c) T is decreasing if and only if $f(E) < T$

For example, (b) shows a criterion for an experience to let the trust level increase. If the trust already has some level *T*, it can only increase when an experience with level *E* at least satisfying $f(E) > T$ occurs; otherwise trust will decrease or stay the same. Another way to use this is to determine directly to which equilibrium trust can go if a given experience level *E* is constantly offered; according to criterion (a) this equilibrium level for trust is *f(E)*. Furthermore, from the monotonicity criteria above it can be derived in the following manner that the equilibrium is always attracting. Suppose T_{eq} is an equilibrium for *E*, and $T < T_{eq}$; this implies

$$T < T_{eq} = f(E)$$

and therefore *T* is increasing for the given *E* by the criterion (b) above. Similarly, when $T > T_{eq}$ for the given *E* it is found that *T* is decreasing by criterion (c). This proves that the process will always converge to the equilibrium, independent of the function *f*. This will also be confirmed by the analysis of the behaviour around the equilibrium below.

Determining the behaviour around an equilibrium

Independent of the precise form of the function f (and hence also independent of the bias parameter β), the behaviour around an equilibrium for a given constant experience E can be found here as follows. Write $T(t) = T_{eq} + \delta(t)$, with $\delta(t)$ the deviation of T from the equilibrium T_{eq} for which it holds $f(E) = T_{eq}$.

$$T(t + \Delta t) = T(t) + \gamma\,(f(E) - T(t))\Delta t$$
$$T_{eq} + \delta(t + \Delta t) = T_{eq} + \delta(t) + \gamma\,\Big(f(E) - \big(T_{eq} + \delta(t)\big)\Big)\Delta t$$
$$\delta(t + \Delta t) = \delta(t) + \gamma\,(f(E) - T_{eq} - \delta(t))\Delta t$$
$$\delta(t + \Delta t) = \delta(t) - \gamma\,\delta(t)\,\Delta t$$
$$\frac{d\delta(t)}{dt} = -\gamma\delta(t)$$

As a differential equation this can be solved analytically using an exponential function:

$$\delta(t) = \delta(0)\, e^{-\gamma t}$$

This shows that the speed of convergence directly relates to parameter γ, and the *convergence rate* defined as reduction factor of the deviation per time unit is

$$cr = e^{-\gamma}$$

This is independent of β, or the function *f*. More specifically, since $\gamma > 0$, the convergence rate is always < 1; from this it follows that the equilibrium is always attracting.

This shows that the speed by which trust adapts to a certain experience level is independent of the specific function f and bias parameter β; it is higher when γ is higher and lower when γ is lower.

5.2 Mathematical analysis of trust models with biased experience also affected by trust

For the models that express the bias based both upon the experience and the current trust level, the following difference equation was used:

$$T(t + \Delta t) = T(t) + \gamma\, f\big(E(t), T(t)\big)\Delta t$$

with $\gamma > 0$. In a similar manner as above the following criteria are obtained:

Equilibrium, increasing and decreasing: biased experience also affected by trust

(a) T is in equilibrium for a given E if and only if $f(E,T) = 0$
(b) T is increasing if and only if $f(E,T) > 0$
(c) T is decreasing if and only if $f(E,T) < 0$

This again shows a criterion, for example, for an experience to let the trust level increase. If the trust already has some level *T*, it can only increase when an experience with level *E* at time *t* at least satisfying $f(E,T) > 0$ is obtained; otherwise trust will decrease or stay the same.

Furthermore, some criterion on the function f can be found in order that the equilibrium T_{eq} for *E* is attracting. Attracting means that if *T* is close to T_{eq} with $T < T_{eq}$, then for the given *E* it should be the case that *T* increases, which according to the above is equivalent with $f(E, T) > 0$. So, starting from $T = T_{eq}$ with $f(E, T_{eq}) = 0$, when *T* is taken lower, the value of $f(E, T)$ has to become higher:

$T < T_{eq} \Rightarrow f(E,T) > f(E,T_{eq})$

This is equivalent with the criterion that in (E, T_{eq}) the function f is decreasing in its second argument: $\partial f/\partial T(E,T_{eq}) < 0$. Below this will be confirmed from the analysis of the behaviour around an equilibrium. This shows that not all functions f will provide the property that the trust levels converge to such an equilibrium value. For a choice to be made for some function f this has to be considered. Below it will be shown that for the choices made in the current paper this criterion is always fulfilled.

Determining the behaviour around an equilibrium
Depending on the form of the function f and also on the bias parameter β, the behaviour around an equilibrium for a given constant experience E can be found as follows. Write $T(t) = T_{eq} + \delta(t)$, with $\delta(t)$ the deviation from the equilibrium T_{eq} for which it holds $f(E,T_{eq}) = 0$. For f the first-order Taylor approximation around T_{eq} in its second argument is used, where $\partial f/\partial T$ denotes the partial derivative of f with respect to its second argument T:

$$f(E,T) = f(E,T_{eq}) + \partial f/\partial T(E,T_{eq})\ (T - T_{eq})$$

Using this it holds

$$f\left(E,T_{eq} + \delta(t)\right) = f(E,T_{eq}) + \partial f/T(E,T_{eq})\,\delta(t) + \partial f/T(E,T_{eq})\,\delta(t)$$
$$f\left(E,T_{eq} + \delta(t)\right) = \partial f/\partial T(E,T_{eq})\ \delta(t)$$

Then the following is obtained:

$$T(t + \Delta t) = T(t) + \gamma\ \left(f(E,T(t))\right)\Delta t$$
$$T_{eq} + \delta(t+\Delta t) = T_{eq} + \delta(t) + \gamma\, f\left(E,T_{eq} + \delta(t)\right)\Delta t$$
$$\delta(t+\Delta t) = \delta(t) + \gamma\left(\partial f/\partial T(E,T_{eq})\right)\delta(t)\Delta t$$
$$d\delta(t)/dt = \gamma\left(\partial f/\partial T(E,T_{eq})\right)\delta(t)$$

As a differential equation this can be solved analytically using an exponential function:

$$\delta(t) = \delta(0)\ e^{-\gamma\,(\partial f/\partial T(E,T_{eq}))t}$$

The convergence rate is defined as reduction factor of the deviation per time unit; this is $e^{-\gamma\,(\partial f/\partial T(E,T_{eq}))}$. This provides a condition on when an equilibrium is attracting, namely $\partial f/\partial T(E,T_{eq}) < 0$. Note that in this case the convergence speed does not only depend on γ but also on f, which in principle relates to the bias β. This speed is higher when γ is higher, but also when $\partial f/\partial T(E,T_{eq})$ is more negative.

5.3 Mathematical analysis of trust models with bias solely determined by current trust

For the models that express the bias based only upon the current trust level, the following difference equation was used:

$$T(t + \Delta t) = T(t) + \gamma\,(E(t) - f(T(t))\,\Delta t$$

where $\gamma > 0$. Similarly the following criteria are found:

Equilibrium, increasing and decreasing:

bias solely determined by current trust

(a) T is in equilibrium for a given E if and only if $E = f(T)$
(b) T is increasing if and only if $E > f(T)$
(c) T is decreasing if and only if $E < f(T)$

Like before, this shows a criterion, for example, for an experience to let the trust level increase. If the trust already has some level T, it can only increase when an experience with level E at least satisfying $E > f(T)$ is obtained; otherwise trust will decrease or stay the same. Moreover, a criterion on the function f can be found in order that the equilibrium T_{eq} for E is attracting. As before note that attracting means that if T is close to with $T < T_{eq}$, then for the given E it should be the case that T increases, which according to criterion (b) above is equivalent with $f(T) < E$. So, starting from $T = T_{eq}$ with $E = f(T_{eq})$, when T is taken lower, the value of $f(T)$ becomes lower:

$$T < T_{eq} \Rightarrow f(T) < f(T_{eq})$$

This means that in T_{eq} the function f has to be increasing: $df/dT(T_{eq}) > 0$. Below, this criterion for being attracting will be confirmed when the behaviour around an equilibrium is analysed. This shows again that not all functions f will provide the property that the trust levels converge to an equilibrium value. For a choice to be made for some function f this criterion $df/dT(T_{eq}) > 0$ has to be taken into account. Below it will be shown that for the choices made in the current paper this criterion is always fulfilled.

Determining the behaviour around an equilibrium
Again, depending on the form of the function f and also on the bias parameter β, the behaviour around an equilibrium for a given constant experience E can be found as follows. Write $T(t) = T_{eq} + \delta(t)$, with $\delta(t)$ the deviation from the equilibrium T_{eq} for which it holds $E = f(T_{eq})$. For f the first-order Taylor approximation around T_{eq} is used:

$$f(T) = f(T_{eq}) + df/dT(T_{eq}) \quad (T - T_{eq})$$

Using this it is obtained:

$$T(t + \Delta t) = T(t) + \gamma \left(E - f(T(t))\right)\Delta t$$
$$T_{eq} + \delta(t + \Delta t) = T_{eq} + \delta(t) + \gamma \Big(E - f\left(T_{eq} + \\ +\delta(t)\right)\Big)\Delta t$$
$$\delta(t + \Delta t) = \delta(t) + \gamma \left(E - f(T_{eq} + \delta(t))\right)\Delta t$$
$$\delta(t + \Delta t) = \delta(t) + \gamma \Big(E - f(T_{eq}) \\ - \left(df/dT(T_{eq})\right)\delta(t)\Big)\Delta t$$
$$\delta(t + \Delta t) = \delta(t) - \gamma\left(df/dT(T_{eq})\right)\delta(t)\Delta t$$
$$d\delta(t)/dt = -\gamma\left(df/dT(T_{eq})\right)\delta(t)$$

As a differential equation this can be solved analytically using an exponential function:

$$\delta(t) = \delta(0)e^{-\gamma(df/dT(T_{eq}))t}$$

This shows that the speed of convergence does not only relate to parameter γ, but also to $df/dT(T_{eq})$ which in principle relates to the bias β. The convergence rate defined as reduction factor of the deviation per time unit is

$$cr = e^{-\gamma\,(df/dT(T_{eq}))}$$

So, also in this case the convergence speed does not only depend on γ but also on f, which in principle relates to the bias β. This speed is higher when γ is higher, but also when $df/dT(T_{eq})$ is higher.

Table 4. Results of the mathematical analysis for the general case

bias depends on	increasing/decreasing	equilibrium value	convergence rate	attracting
only on experience	$f(E) > T$, $f(E) < T$	$f(E) = T_{eq}$	$e^{-\gamma}$	*always*
on experience and trust	$f(E,T) > 0$, $f(E,T) < 0$	$f(E,T_{eq}) = 0$	$e^{\gamma\,(\partial f/\partial T(E,T_{eq}))}$	$\partial f/\partial T(E,\ T_{eq}) < 0$
only on trust	$E > f(T)$, $E < f(T)$	$E = f(T_{eq})$	$e^{-\gamma\,(df/dT(T_{eq}))}$	$df/dT(T_{eq}) > 0$

5.4 Mathematical analysis of the example biased trust models for the three types

In this section, for each of the three general types of biased trust models analysed above, it will be investigated how the criteria can be formulated more specifically for the linear functions used in the current paper as instances for the function f: LiE, LiET, and LiT,

5.4.1 More specific analysis for the linear case of bias only depending on experience (LiE)

For the first case the following linear function was addressed (**LiE**):

$f(E) = E + (2\beta - 1)(1 - E)$ when $\beta \geq 0.5$
$f(E) = 2\beta E$ when $\beta \leq 0.5$

Case $\beta \geq 0.5$

Criterion for increasing for LiE with $\beta \geq 0.5$
$E + (2\beta - 1)(1 - E) > T$
$E + (2\beta - 1) - (2\beta - 1)E > T$
$2(1 - \beta)E + (2\beta - 1) > T$
$2(1 - \beta)E > T - (2\beta - 1)$
$E > (T(t) - (2\beta - 1))/\big(2(1 - \beta)\big)$
$E > (T(t) - 1 - (2\beta - 2))/(2(1 - \beta))$
$E > (T(t) - 1)/(2(1 - \beta)) - \big(2(\beta - 1)\big)/\big(2(1 - \beta)\big)$
$E > 1 - \frac{1}{2}(1 - T)/(1 - \beta)$

Criterion for decreasing for LiE with $\beta \geq 0.5$
$E < 1 - \frac{1}{2}(1 - T)/(1 - \beta)$

Criterion for equilibrium for LiE with $\beta \geq 0.5$
$E = 1 - \frac{1}{2}(1 - T)/(1 - \beta)$
$E(1 - \beta) = (1 - \beta) - \frac{1}{2}(1 - T)$
$\frac{1}{2}(1 - T) = (1 - \beta) - E\,(1 - \beta)$
$1 - T = 2(1 - \beta) - 2E\,(1 - \beta)$
$T = 1 - 2\,(1 - \beta)(1 - E)$

Note that for $\beta = 0.5$ (no bias) the criterion for an equilibrium is $E = T$, what is to be expected. For $\beta = 0.75$, the criterion is

$E = 1 - \frac{1}{2}(1 - T)/0.25$
$E = 1 - 2(1 - T)$
$E = 1 - 2\,(1 - T)$
$E = 2T - 1$

Note that for lower values of T this can provide a negative number. However, as the experience cannot be lower than 0, this implies that for such values of T no equilibrium occurs. For $\beta = 0.875$, the criterion is

$E = 1 - \frac{1}{2}(1 - T)/0.125$
$E = 1 - 4\,(1 - T)$
$E = 4T - 3$

For β approaching 1, the criterion always becomes a negative number (implying increase), unless $T = 1$; this implies that for this value of β no equilibrium occurs except for $T = 1$ and any value for E.

Behaviour around the equilibrium for LiE with $\beta \geq 0.5$

For this case the behaviour around the equilibrium does not depend on the specific form of the function f. The convergence rate is: $cr = e^{-\gamma}$, which is independent, for example, of β. As $\gamma > 0$, the equilibrium is always attracting.

Case $\beta \leq 0.5$

Criterion for increasing for LiE with $\beta \leq 0.5$

$2\beta E > T$
$E > \frac{1}{2}T/\beta$

Criterion for decreasing for LiE with $\beta \leq 0.5$

$E < \frac{1}{2}T/\beta$

Criterion for equilibrium for LiE with $\beta \leq 0.5$

$E = \frac{1}{2}T/\beta$
$T = 2\beta E$

Behaviour around the equilibrium for LiE with $\beta \leq 0.5$

For this case the behaviour around the equilibrium does not depend on the specific form of the function f. The convergence rate is: $cr = e^{-\gamma}$, which is independent, for example, of β or E. As $\gamma > 0$, the equilibrium is always attracting.

5.4.2 More specific analysis for the linear case of bias depending on both experience and trust (LiET)

For the second case the following linear function was addressed (**LiET**):

$$f(E,T) = \beta\big(1 - (1 - E)(1 - T)\big) + \big(1 - \beta\big)E\,T - T$$

For the linear example the inequalities and equation can be explicitly solved as follows.

Criterion for increasing for LiET

$\beta\,\big(1 - (1 - E)(1 - T)\big) + (1 - \beta)ET - T > 0$
$\beta\,\big(1 - (1 - T) + E\,(1 - T)\big) + (1 - \beta)ET - T > 0$
$\beta\,T + E\,\beta\,(1 - T) + (1 - \beta)ET - T > 0$
$E\,(\beta\,(1 - T) + (1 - \beta)T) > (1 - \beta)T$
$E\,(\beta - \beta T) + T - \beta T) > (1 - \beta)\,T$
$E\,\big(\beta + (1 - 2\beta)T(t)\big) > (1 - \beta)T$
$E > ((1 - \beta)\,T)(1 - \beta)\,T(t)\,/(\beta + (1 - 2\beta)T)$

Criterion for decreasing for LiET

$E < (1 - \beta)\,T/(\beta + (1 - 2\beta)T)$

Criterion for equilibrium for LiET

$E = (1 - \beta)\,T/(\beta + (1 - 2\beta)T)$
$E(\beta + (1 - 2\beta)T) = (1 - \beta)\,T$
$E\beta + (1 - 2\beta)ET = (1 - \beta)\,T$
$(1 - \beta)T - (1 - 2\beta)ET = E\,\beta$
$\big((1 - \beta) - (1 - 2\beta)E\big)T = E\,\beta$
$T = E\beta\,/\big((1 - \beta) - (1 - 2\beta)E(t)\big)$

Behaviour around the equilibrium for LiET

For the specific linear function f used above, it holds:

$$\frac{\partial f}{\partial T(E,T)} = \frac{\partial\big(\beta\big(1 - (1 - E)(1 - T)\big) + (1 - \beta)ET - T\big)}{\partial T(E,T)}$$
$= \beta(1 - E) + (1 - \beta)E - 1$

Using this, for the linear case it is obtained:
$\delta(t) = \delta(0)\,e^{\gamma\,(\beta(1-E)+(1-\beta)E-1)}$

and the convergence rate is $e^{\gamma\,(\beta(1-E)+(1-\beta)E-1)}$. This shows that for this case the speed of convergence not only relates to parameter γ, but also to β and E. More specifically, the convergence rate is < 1 if and only if

$\beta\,(1 - E) + (1 - \beta)E - 1 < 0$

This is a condition for an equilibrium to be attracting. It can be rewritten into an explicit criterion for E as follows:

$\beta - \beta E + E - \beta E - 1 < 0$
$1 - \beta - E + \beta E + \beta E > 0$
$(1 - \beta\,)(1 - E\,) + \beta E > 0$

This is always the case.

5.4.3 More specific analysis for the case of bias depending only on trust (LiT)
For the third case the following function was addressed (**LiT**):

$f(T) = T - (1-T)\,(2\beta - 1)(1-T)$ when $\beta \geq 0.5$
$f(T) = T - (1-T)(2\beta - 1)T$ when $\beta \leq 0.5$

This can be analysed more specifically as follows

Case $\beta \geq 0.5$

Criterion for increasing for LiT with $\beta \geq 0.5$
$E > f(T) = T - (1-T)\,(2\beta - 1)\,(1 - T)$

Criterion for decreasing for LiT with $\beta \geq 0.5$
$E < f(T) = T - (1-T)\,(2\beta - 1)\,(1 - T)$

Criterion for equilibrium for LiT with $\beta \geq 0.5$
$E = f(T) = T - (1-T)(2\beta - 1)(1-T)$
$E = T - (2\beta - 1)(T^2 - 2T + 1)$
$E = T - (2\beta - 1)T^2 + 2(2\beta - 1)T - (2\beta - 1)$
$E = -(2\beta - 1)T^2 + (4\beta - 1)T - (2\beta - 1)$
$(2\beta - 1)T^2 - (4\beta - 1)T + (2\beta - 1) + E = 0$

For the special case that $\beta = 0.5$ (no bias) this latter criterion reduces to a linear equation $-T + E = 0$ with solution $T = E$. For the general case $\beta > 0.5$ the above expression is a quadratic equation in T with discriminant

$D = (4\beta - 1)^2 - 4(2\beta - 1)\big((2\beta - 1) + E\big)$
$= (16\beta^2 - 8\beta + 1) - 4(4\beta^2 - 4\beta + 1) - 4(2\beta - 1)E$
$= 16\beta^2 - 8\beta + 1 - 16\beta^2 + 16\beta - 4 - 4(2\beta - 1)E$
$= 8\beta - 3 - 4(2\beta - 1)E$
$= 8\beta - 3 - (8\beta - 4)E$
$= 8(1 - E)\beta + 4E - 3$

From this expression for D, which is linear in both β and E, given that $\beta \geq 0.5$ it can easily be seen that D is always ≥ 1:

- for $\beta = 0.5$ it holds $D = 4(1 - E) + 4E - 3 = 1$
- for $\beta = 1$ it holds $D = 8(1 - E) + 4E - 3 = 5 - 4E \geq 1$ since $E \leq 1$

Alternatively, considering special values of E:

- for $E = 1$ it holds $D = 1$
- for $E = 0$ it holds $D = 8\beta - 3 \geq 4 - 3 = 1$ since $\beta \geq 0.5$

Therefore D is positive and the quadratic equation has two solutions for T

$T_{1,2} = \big((4\beta - 1) \pm \sqrt{D}/\big(2(2\beta - 1)\big)\big)$
$= \Big((4\beta - 1) \pm \sqrt{(8(1 - E)\beta + 4E - 3)}/\big(2(2\beta - 1)\big)\Big)$

Since $D \geq 1$ for the highest solution T_2 it holds

$T_2 \geq \big((4\beta - 1) + 1\big)/\big(2(2\beta - 1)\big) = 4\beta/(4\beta - 2)$
$= (4\beta - 2)/(4\beta - 2) + 2/(4\beta - 2)$

$= 1 + 2/(4\beta - 2) > 1$

Similarly, from $D \geq 1$ it follows that for the lowest solution T_1 (for the -) it holds

$T_1 \leq ((4\beta - 1) - 1)/2(2\beta - 1) = (4\beta - 2)/2(2\beta - 1)$

$= 1$

Therefore the equilibrium T_{eq} for a given E is the lowest solution T_1

$T_{eq} = \frac{(4\beta - 1) - \sqrt{D}}{2(2\beta(1)} = \frac{(4\beta - 1) - \sqrt{[8(1 - E)\,\beta + 4E - 3]}}{2(2\beta - 1)} \leq 1$

This is a positive number since $\sqrt{D} \leq (4\beta - 1)$ as can be seen from the initial expression
$D = (4\beta - 1)^2 - 4(2\beta - 1)\big((2\beta - 1) + E\big)$
$\leq (4\beta - 1)^2$

Behaviour around the equilibrium for LiT with $\beta \geq 0.5$
It holds

$df/dT(T) = 1 + 2(1 - T)(2\beta - 1)$
$df/dT\big(T_{eq}\big) = 1 + 2\big(1 - T_{eq}\big)(2\beta - 1)$

Therefore for this case the convergence rate is

$cr = e^{-\gamma(df/\boldsymbol{d}T(Teq))} = e^{-\gamma\left(1 + 2\left(1 - T_{eq}\right)(2\beta - 1)\right)}$
This depends both on γ and β, and via T_{eq} also on E.

The criterion for the equilibrium being attracting is that $\boldsymbol{d}f/\boldsymbol{d}T(T_{eq}) > 0$. This is equivalent to:

$1 + 2\big(1 - T_{eq}\big)(2\beta - 1) > 0$

As $\beta \geq 0.5$, this is always the case.

Case $\beta \leq 0.5$

Criterion for increasing for LiT with $\beta \leq 0.5$
$E > T + (1 - T)\,(T - 2\,\beta\,T)$
$= 2(1 - \beta)T - T^2(1 - 2\,\beta)$

Criterion for decreasing for LiT with $\beta \leq 0.5$

$$E < T + (1-T)\big(T - 2\,\beta\,T\big)$$
$$= 2(1-\beta)T - T^2(1-2\,\beta)$$

Criterion for equilibrium for LiT with $\beta \leq 0.5$

$E = T + (1-T)\,(T - 2\,\beta\,T)$
$E = 2(1-\beta)T - T^2(1-2\,\beta)$
$(1-2\beta)T^2 - 2(1-\beta)T + E = 0$

This is a quadratic equation in T with discriminant

$D = 4(1-\beta)^2 - 4\,(1-2\beta)\,E$
Then
$T_{1,2} = \ (2(1-\beta) \pm \sqrt{D})/2(1-2\beta)$
$= \left((1-\beta) \pm \sqrt{(1-\beta)^2 - (1-2\beta)\,E}\right) / (1-2\beta)$

Solutions for T require that $D \geq 0$; this is equivalent to:

$(1-\beta)^2 - (1-2\beta)E \geq 0$
$(1-\beta)^2 \geq (1-2\beta)E$
$E \leq (1-\beta)^2/(1-2\beta)$
$E \leq 1 + \beta^2/(1-2\beta)$

As $E \leq 1$ and $1 + \beta^2/(1-2\beta) > 1$, this is always fulfilled. The highest solution T_2 is > 1 as can be seen from

T_2
$= \left((1-\beta) + \sqrt{((1-\beta)^2 - (1-2\beta)E)}\right)/(1-2\beta)$
$\geq (1-\beta)\ /\ (1-2\beta)$
$= (1-2\beta)\ /\ (1-2\beta) + \beta\ /\ (1-2\beta)$
$= 1\ +\ \beta/(1-2\beta) > 1$

Therefore the equilibrium value T_{eq} is the smallest solution T_1

$$T_{eq} = \left((1-\beta) - \sqrt{(1-\beta)^2 - (1-2\beta)\,E}\right)/(1-2\beta)$$

As above it can be seen that this is a positive number.

Behaviour around the equilibrium for LiT with $\beta \leq 0.5$

It holds

$\boldsymbol{df/dT}(T) = 1 - (1-T)\,(2\beta-1) + (2\beta-1)T$
$=\ 1 - (2\beta-1) + 2(2\beta-1)T$
$=\ 2(1-\beta) + 2(2\beta-1)T$

$\boldsymbol{df/dT}(T_{eq}) = 2(1-\beta)\ +\ 2(2\beta-1)\,T_{eq}$

Therefore for this case the convergence rate is

$cr = e^{-\gamma(df/dT(T_{eq}))} = e^{-\gamma(2(1-\beta)\,+\,2(2\beta-1)\,T_{eq})}$

This depends both on γ and β, and via T_{eq} also on E. The criterion for the equilibrium being attracting is that $df/dT(T_{eq}) > 0$. This is equivalent to:

$2(1-\beta) + 2(2\beta-1)T_{eq} > 0$

As $\beta \leq 0.5$, this is always the case.

Table 5. Results of the mathematical analysis for the specific linear functions

	bias depends on	increasing/decreasing	equilibrium value	convergence rate	Attracting
LiE	only on experience: $\beta \geq 0.5$	$E > 1 - ½\,(1\ -\ T)/(1-\beta)$ $E < 1 - ½\,(1\ -\ T)/(1-\beta)$	$E = 1 - ½(1-T_{eq})/(1-\beta)$ $T_{eq} = 1 - 2\,(1-\beta)(1-E)$	$e^{-\gamma}$	*Always*
	only on experience: $\beta \leq 0.5$	$E > ½T/\,\beta$ $E < ½T/\,\beta$	$E = ½\,T_{eq}\,/\,\beta$ $T_{eq} = 2\beta E$	$e^{-\gamma}$	*Always*
LiET	on experience and trust	$E > (1-\beta)T/(\beta + (1-2\beta)T)$ $E < (1-\beta)T/(\beta + (1-2\beta)T)$	$E = (1-\beta)T_{eq}/(\beta + (1-2\beta)T_{eq})$ $T_{eq} = E\beta/((1-\beta) - (1-2\beta)E)$	$e^{\gamma(\beta\,(1-E)+(1-\beta)E-1)}$	*Always*
LiT	only on trust: $\beta \geq 0.5$	$E > T - (1-T)(2\beta-1)(1-T)$ $E < T - (1-T)(2\beta-1)(1-T)$	$E = T_{eq} - (1-T_{eq})(2\beta-1)(1-T_{eq})$ $T_{eq} = \frac{(4\beta-1) - \sqrt{[8(1-E)\,\beta + 4E - 3]}}{2(2\beta\ -1)}$	$e^{-\gamma\left(1+2(1-T_{eq})(2\beta-1)\right)}$	*Always*
	only on trust: $\beta \leq 0.5$	$E > 2(1-\beta)T - T^2(1-2\,\beta)$ $E < 2(1-\beta)T - T^2(1-2\,\beta)$	$E = 2(1-\beta)T_{eq} - (T_{eq})^2(1-2\beta)$ $T_{eq} = \frac{(1-\beta) - \sqrt{[(1-\beta)^2 - (1-2\beta)\,E}}{1-2\beta}$	$e^{-\gamma\left(2(1-\beta) - 2(1-2\beta)T_{eq}\right)}$	*Always*

6. Human-Based Trust Experiment

In this section the human-based trust experiment is explained. In Section 6.1 the participants are described. In Section 6.2 an overview of the used experimental environment is given. Thereafter, the procedure of the experiment and data collection is explained in Sections 6.3.

6.1 Participants

Eighteen participants (eight male and ten female) with an average age of *23 (SD = 3.8)* participated in the experiment as paid volunteers. Non-colour blinded participants were selected. All were experienced computer users, with an average of *16.2* hours of computer usage each week *(SD = 9.32)*.

6.2 Task

As the bias-based trust models are designed to work in situations in which humans have to decide to trust either one of multiple heterogeneous trustees, the experimental task used involved three different trustees, namely two human participants and a support system. The task was a classification task in which the two participants on two separate personal computers had to classify geographical areas according to specific criteria as areas that either needed to be attacked, helped or left alone by ground troops. 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 be 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 the participants could update their internal trust models for each trustee (self, other, software agent).

In Fig. 6 the interface of the task is shown. The map is divided in *10 x 10* areas. These boxes are the areas that were classified. The first UAV starts in the top left corner and the second one left in the middle. The UAVs fly a predefined route so participants do not have to pay attention to navigation. The camera footage of the upper UAV is positioned top right and the other one bottom right.

The advice of the self, other and the software agent was communicated via dedicated boxes below the camera images. The advice to attack, help, or do nothing was communicated by red, green and yellow, respectively. On the overview screen on the left, feedback was communicated by the appearance of a green tick or a red cross. The reliance decision of the participant is also shown on the overview screen behind the feedback (feedback only shown in the feedback phase). The phase depicted in Figure 6 was the reliance phase before the participant indicated his reliance decision.

6.3 Data Collection

During the above described experiment, input and output were logged using a client-server application. The interface of this application is shown in Fig. 7. Two other client machines, that were responsible for

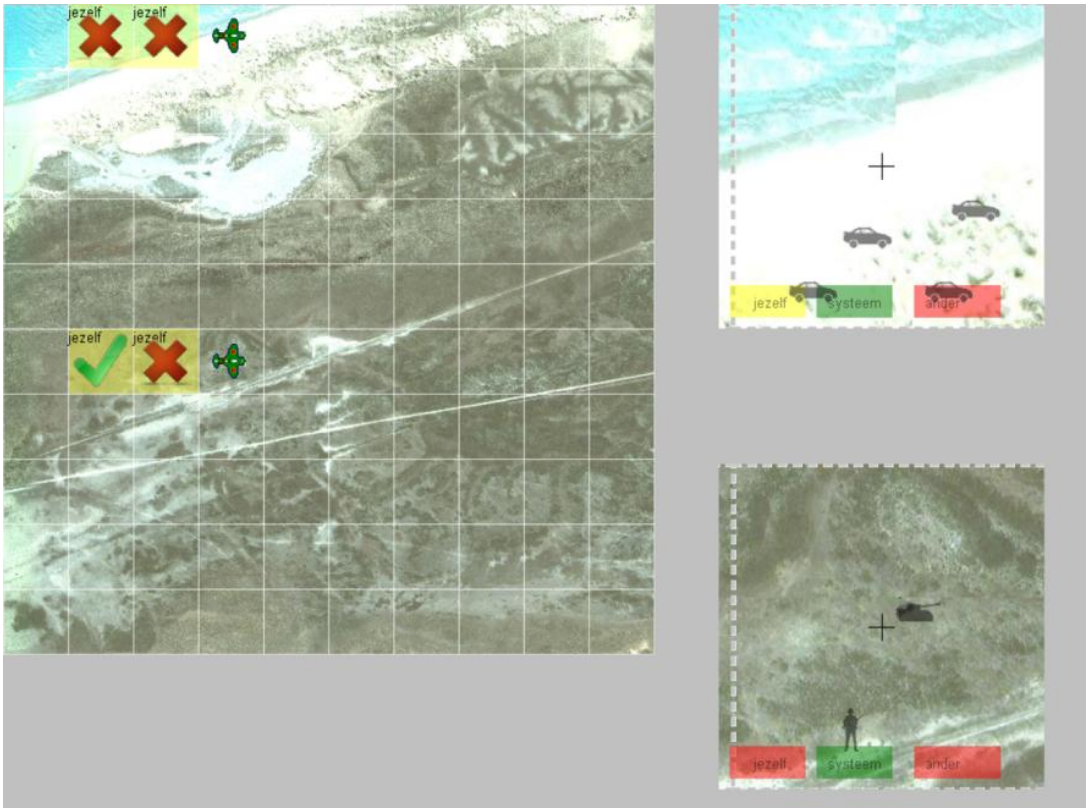


Fig. 6. Interface of the task

executing the task as described in the previous subsection, were able to connect via a local area network to the server, which was responsible for logging all data and communication between the clients. The interface shown in Fig. 7 could be used to set the client's IP-addresses and ports, as well as several experimental settings, such as how to log the data. In total the experiment lasted approximately *15* minutes per participant.

Experienced performance feedback of each trustee and reliance decisions of each participant were logged in temporal order for later analysis. During the feedback phase the given feedback was translated to a penalty of either *0*, *0.5* or *1*, representing a good, neutral or poor experience of performance, respectively. This directly maps to the value $(E_i(t)+1) / 2$ in the trust models. During the reliance phase the reliance decisions were translated to either *0* or *1* for each trustee S_i, which represented that one relied or did not rely on S_i.

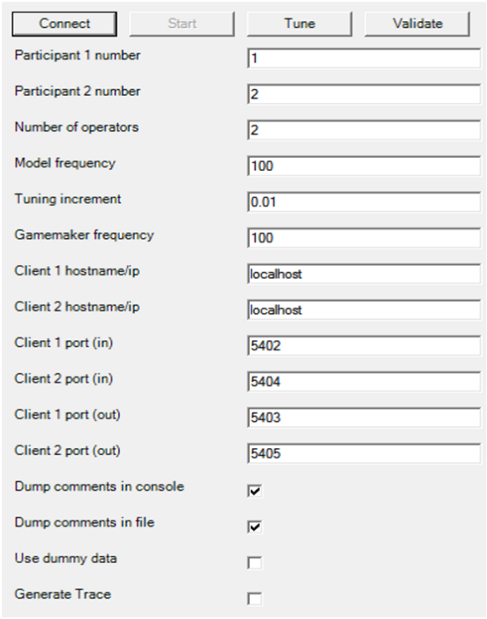


Fig. 7. Interface of the application used for gathering validation data (Connect), for parameter adaptation (Tune) and validation of the trust models (Validate).

7. Validation of Bias-based Trust Models

In this section validation process of the trust models described in section 2 are presented. In section 7.1 parameter adaption technique is explained, 7.2 and 7.3, explain the model validation process and results for bias-based trust models, respectively.

7.1 Parameter Adaptation

The data collection described in Section 6.3 was repeated twice on each group of two participants, called condition 1 and condition 2, respectively. The data from one of the conditions was used for parameter adaptation purposes for each model, and the data from the other condition for model validation (see Section 6.3). This process of parameter adaptation and validation was balanced over conditions, which means that condition 1 and condition 2 switch roles, so condition 1 is initially used for parameter adaptation and condition 2 for model validation, and thereafter condition 2 is used for parameter adaptation and condition 1 for model validation (i.e. cross-validation). Then the average was calculated of the two calculated validities, per participant, per model. This last value is called the accuracy of the models. The results are in the form of accuracies per trust model and their differences are detected using a repeated measures analysis of variance (ANOVA) and post-hoc Bonferroni t-tests.

After the different models were tuned, the best fit model (with the maximum accuracy) is selected based on the maximum accuracy for the participant at hand. This was done because at the moment one does not know beforehand which bias type will be suitable for the specific participant. 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)).

Both the parameter adaptation and model validation procedure was done using the same application as was used for gathering the empirical data. The interface shown in Figure 7 could also be used to alter validation and adaptation settings, such as the granularity of the adaptation.

The number of parameters of the models presented in Section 2 to be adapted for each model and each participant suggest that an exhaustive search as described in [6] for the optimal parameter values is feasible. This means that the entire parameter search space is explored to find a vector of parameter settings resulting in the maximum accuracy (i.e., the amount of overlap between the model's predicted reliance decisions and the actual human reliance decisions) for each of the models and each participant. The corresponding code of the implemented exhaustive search method is shown in Algorithm 1.

ALGORITHM 1: ES-PARAMETER-ADAPTATION(E, R_H)

```
1  δ_best = ∞, X = 0
2  for all parameters x in vector X do
3    for all settings of x do
4      δ_x = 0
5      for all time points t do
6        e = E(t), r_M = R_M(e, X), r_H = R_H(e)
7        if r_M not equal r_H then
8          δ_x = δ_x+1
9        end if
10     end for
11     if δ_x < δ_best then
12       X_best = X, δ_best = δ_x
13     end if
14   end for
15 end for
16 return X_best
```

In this algorithm, $E(t)$ is the set of experiences (i.e., performance feedback) at time point t for all trustees, $R_H(e)$ is the actual reliance decision the participant made (on either one of the trustees) given a certain experience e, $R_M(e,X)$ is the predicted reliance decision of the trust model M, given an experience e and candidate parameter vector X (reliance on either one of the trustees), X is the distance between the estimated and actual reliance decisions given a certain candidate parameter vector X, and δ_{best} is the distance resulting from the best parameter vector X_{best} found so far. The best parameter vector X_{best} is returned when the algorithm finishes. This parameter adaptation procedure was implemented in Microsoft ® C#.Net 2005 development environment.

In order to compare the different bias-based trust 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 as described in section 6. It is hereby assumed that the human always consults the most trusted trustee. 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.

7.2 Computational Complexity

As the models described in Section 2 have a different number of parameters the parameter tuning process took a different amount of time for each of the 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), while linear models have four ($\gamma, \beta_1, \beta_2, \beta_3$ 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 ($\gamma, \tau_1, \tau_2, \tau_3, \sigma_1, \sigma_2$, and σ_3, where γ and τ has precision *0.01* and σ has precision *1* within range *1 to 20*).

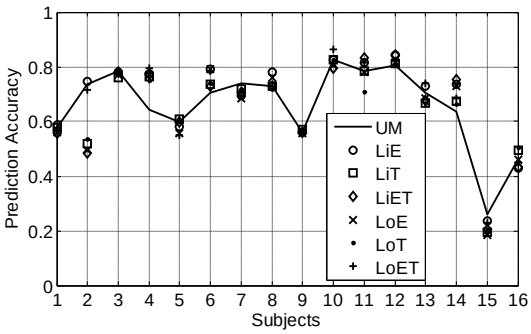
If the time required is calculated say for LoT for tuning one subject then it has *S=1*, *M=1*, *B=1*, *P=7* (4 parameters with precision *0.01*, and *3* parameter with precision *1* and in range (*1-20*), *T=100*3* (to calculate trust value at each time point, predict reliance decision and calculate distance from empirical data), and *N=3* then it counts to $1x1x1x10^{4x2}x20^{3}x3*10^{2}x3=7.2x10^{14}$, which on *2.4* MHz computer will take approx. *3.47* days. For a linear model it counts to *37.5* seconds. So to validate all seven models against one subject will take *10.41* days. If all subjects are validated for all seven models in a serial fashion (one by one) on a machine having speed *2.4* MHz then it will cost *166.66* days to complete. Hence during the process of tuning two approaches are followed a) to decrease granularity of the parameters from *0.01* to *0.025* (for α, τ, β) and *1 to 2* (for σ) and secondly to use DAS-4 [1] (distributed ASCI super computer version 4) which can distribute validation of each of the subject on separate machine in a distributed cluster. Hence *16* machines on DAS-4 have been utilized for this purpose. On average these machines have provided *0.31* MHz of computing power. These steps have speedup the process very much and the whole process took approximately *6.19* hours at DAS-4 with these parameters.

7.3 Validation Results

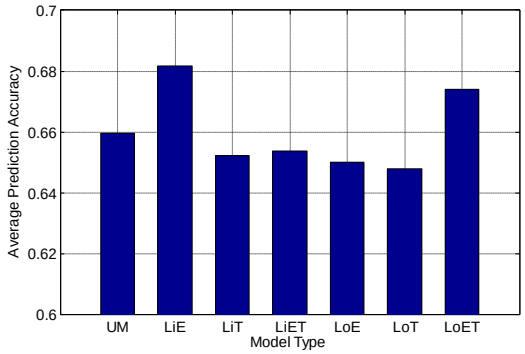
From the data of 18 participants, two outliers have been removed, which leaves a data set of 16 accuracies per model type (UM, LiE, LiT, LiET, LoE, LoT, LoET and MAX).

The actual found tuned parameters per model type per participant are too numerous to show in the paper. Hence we only show the found accuracies.

In Figure 8a 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 Figure 8b 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).



a)



b)

Fig. 8: a) prediction accuracy of models across subjects, b) average prediction accuracy of models for all subjects

In Fig. 9 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.

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, also for unknown data.

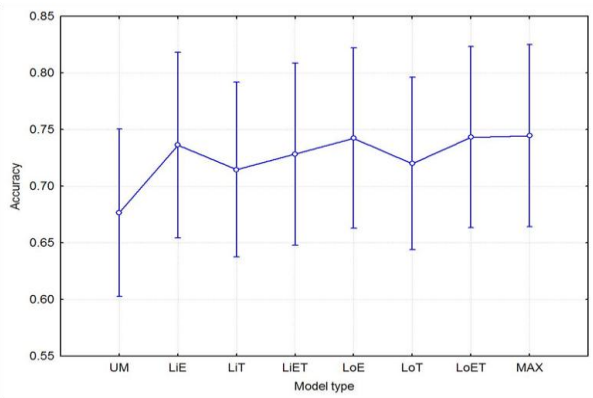


Fig. 9. Main effect of model type on accuracy

8. Discussion and Conclusions

In this paper, approaches have been presented that allow for modelling biases in human trust dynamics. In order to come to models incorporating such approaches, an existing model [11], which is often applied (e.g., [17], [18], [19]), has been extended with additional constructs. 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
(3) a model that uses the previous trust value on which the bias is applied.

Simulation results of the behaviour of each of the model have been shown, as well as a comparison of the behaviour of the models via the mutual model mirroring method presented in [9]. Furthermore, the resulting simulation traces have been formally analysed by means of the verification of formal desired properties and were shown to behave as expected. In addition a detailed mathematical analysis has been performed to investigate dynamic properties of bias-based trust models. The properties addressed include aspects such as when trust is increasing or decreasing, which equilibria are possible (i.e., $T(t + \Delta t) = T(t)$), and how the behaviour of the models is near the equilibria, in particular whether they are attracting and what the rate of convergence to such an equilibrium is. The main goal of the research presented here is to model and validate human bias-based trust. Therefore, an extensive validation has taken place in which the bias-based trust models were used to describe and forecast human trust levels. In this paper, to tailor the model to a specific human, a simple parameter estimation technique has been used, but more complex estimation techniques could also be applied. The tuning technique used for the personalization of trust models was inspired by the techniques presented in [6]. The technique applied being exhaustive in nature consumes a lot of computation power. Hence, during the process of tuning two approaches are followed a) to decrease granularity of the parameters from and secondly to use DAS-4 [1] which can distribute validation of each of the subject on separate machine in a distributed cluster. Hence *16* machines on DAS-4 have been utilized for this purpose. These steps have speedup the process significantly which approximately *6.19* hours at DAS-4 instead of *166* days on a personal computer.

The validation study of bias-based trust models 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.

Within the domain of agent systems, quite some trust models have been developed, see e.g. [13], [14] for an overview. Although the focus of this paper has been on the design of bias-based trust models and validation of these models, other trust models can

also be validated using the experimental data obtained in combination with parameter estimation. This is part of future work. Furthermore, other parameter adaptation methods will be explored or extended for the purpose of real-time adaptation. In addition, a personal assistant software agent will be implemented that is able to monitor and balance the functional state of the human in a timely and knowledgeable manner. Also applications in different domains are explorable, such as the military and air traffic control domain.

In future, given the approach presented in this paper, other models that represent human trust from the literature, for example, addressing trust in agents as teammates (see e.g. [13], [14], [3a]) could also be extended with the notion of human biases. Furthermore it could be investigated that how far these extensions improve the accuracy of those models.

References

[1] Bal, H., Bhoedjang, R., Hofman, R., Jacobs, C., Kielmann, T., Maassen, J., et al. The distributed ASCI Supercomputer project. SIGOPS Operating System Review, vol. 34, no. 4, pp. 76-96, 2000.

[2] Bosse, T., Jonker, C., Meij, L. v. d., Sharpanskykh, A., and Treur, J.. Specification and verification of dynamics in agent models. International Journal of Cooperative Information Systems, vol. 18, pp. 167-193, 2009.

[3] Falcone, R. and Castelfranchi, C. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 740-747, 2004.

[3a] Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., and Endsley, M. R. The Influence of Agent Reliability on Trust in Human-Agent Collaboration. In Proccedings of the European Conference on Cognitive Ergonomics, 2008.

[4] Huff, L. and Kelley, L., (2003). Levels of Organizational Trust in Individualist versus Collectivist Societies: A Seven Nation Study. Organizational Science, vol. 14, pp. 81-90.

[5] Hoogendoorn, M., Jaffry, S.W., Maanen, P.-P. van, and Treur, J., Modeling and Validation of Biased Human Trust. In Boissier, O., et al. (eds.), Proceedings of the Eleventh IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 256-263, 2011, IEEE Computer Society Press.

[6] Hoogendoorn, M., Jaffry, S.W. and Treur, J. Modeling dynamics of relative trust of competitive information agents. In M. Klusch, M. Pechoucek, and A. Polleres, editors, Proceedings of the Twelfth International Workshop on Cooperative Information Agents, Lecture Notes in Artificial Intelligence, vol. 5180, pp. 55-70, 2008, Springer Verlag.

[7] 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.), Proceedings of the Ninth IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 458-465, 2009, IEEE Computer Society Press.

[8] Hoogendoorn, M., Jaffry, S.W., and Treur, J., Cognitive and Neural Modeling of Dynamics of Trust in Competitive Trustees. *Cognitive Systems Research*, 2012, in press.

[9] Jaffry, S.W., and Treur, J. Comparing a Cognitive and a Neural Model for Relative Trust Dynamics. In Leung, C.S., Lee, M., and Chan, J.H. (eds.), Proceedings of Sixteenth International Conference on Neural Information Processing, Part I. Lecture Notes in Computer Science, vol. 5863, pp. 72-83, 2009, Springer Verlag.

[10] Jonker, C. M., Schalken, J. J. P., Theeuwes, J. and Treur, J. Human experiments in trust dynamics. In Proceedings of the Second International Conference on Trust Management, Lecture Notes in Computer Science, vol. 2995, pp. 206-220. 2004, Springer Verlag.

[11] Jonker, C. M. and Treur, J. Formal analysis of models for the dynamics of trust based on experiences. In F.J. Garijo and M. Boman, (eds.), Multi-Agent System Engineering, Proceedings of the Ninth European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Lecture Notes in Computer Science, vol. 1647, pp. 221-232, 1998, Springer Verlag.

[12] Maanen, P.-P. v., Klos, T. and Dongen, K. v. Aiding Human Reliance Decision Making using Computational Models of Trust. In Proceedings of the Workshop on Communication between Human and Artificial Agents, pp. 372-376, 2007, IEEE Computer Society Press.

[13] Ramchurn, S., Huynh, D. and Jennings, N. Trust in multi-agent systems. The Knowledge Engineering Review, vol. 19, pp.1-25, 2004.

[14] Sabater, J. and Sierra, C. Review on Computational Trust and Reputation Models. Artificial Intelligence Review, vol. 24, pp. 33-60, 2005.

[15] Sears, D.O. The Person Positivity Bias. Journal of Personality and Social Psychology, vol. 44, pp. 233-250, 2007.

[16] Sharpanskykh, A., and Treur, J., A Temporal Trace Language for Formal Modelling and Analysis of Agent Systems. In Dastani, M., Hindriks, K.V., and Meyer, J.J.Ch. (eds.), Specification and Verification of Multi-Agent Systems, pp. 317-352, 2010, Springer Verlag.

[17] Singh, S.I., and Sinha, S.K. A New Trust Model Based on Time Series Prediction and Markov Model. In Das, V.V., and Vijaykumar, R. (eds.), Proceedings of the International Conference on Information and Communication Technologies, Communications in Computer and Information Science, vol. 101, pp. 148-156, 2010, Springer Verlag.

[18] Skopik, F., Schall, D. and Dustdar, S. Modeling and Mining of Dynamic Trust in Complex Service-Oriented Systems, Information Systems, vol. 35, pp. 735–757, 2010.

[19] Walter, F.E., Battiston, S. and Schweitzer. F. 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, pp. 197-204, 2009, ACM Press.

[20] 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, 1998.