

# Computationally Modelling Trust: An Exploration

Judith Masthoff

University of Aberdeen, Aberdeen, Scotland, UK  
jmasthoff@csd.abdn.ac.uk

Trust is a popular and much disputed topic in various research communities. This paper attempts to integrate existing knowledge on trust into a simple computational model. The model incorporates the impact of direct experiences, reputation, stereotypes, empathy and user characteristics on trust. We present the results of two exploratory experiments testing and improving aspects of the model.

## 1 Introduction

Trust has been extensively studied in fields as varied as economics [13,17], business and marketing [8,24], politics, e-commerce [1,14,41], psychology [7,42,50], sociology [27], medicine, nursing [15], and computing science [10,11,19,39,40,44]. Whilst many definitions of trust have been proposed, none has been agreed upon. However, there have been some good efforts in inventorizing the different definitions, and trying to distil what characterizes trust and what different dimensions there are to it [15,29]. As noted by [29], different researchers have defined different types of trust. Based on their review, [29] defined six types of trust. Trust in this paper is closest to the one they called Trusting Beliefs, defined in [29] as “the extent to which one believes (and feels confident in believing) that the other person is trustworthy in the situation”, with trustworthy defined as “willing and able to act in the other person’s best interests”. The computational model provided in this paper will make our view on trust more explicit.

Trust is key to long-term relationships [22], and user trust has been shown to affect the success of a system, in terms of increasing sales, users’ likelihood to stay with and return to the system [8], use of information provided [32], and users’ willingness to pay more [1]. A computational model of trust is particularly useful when an adaptive system interacts with a community of users (like a group recommender system does). Simply optimizing trust is then not possible: actions aimed to increase the trust of some users may well decrease the trust of others. An accurate trust model would allow tailoring of system actions (such as selected appearance for an embodied agent, empathetic explanations, items) to maintain the trust of all users. In a social system, models of users’ trust can also be used when facilitating interactions between users.

Many factors influence the trust of an agent  $x$  in another agent  $y$ . In this paper, we will consider the following: (1) direct experiences of  $x$  with  $y$ , (2) indirect experiences, namely the reported experiences of others (often referred to as reputation [40]), (3) intuitions of  $x$  about  $y$  based on stereotypes, (4) empathy between  $x$  and  $y$ , and (5) characteristics of  $x$ . Sections 2 and 4 discuss these factors, and incorporate them into a simple computational model of trust. Sections 3 and 5 report on two experiments that investigate some of the issues arising from the modelling and propose improvements. Section 6 concludes this paper.

## 2 Modelling direct experience and reputation

### 2.1 Direct Experience. “You have seen me act well”

We will denote the impact of direct experience  $d$  on the trust of  $x$  in  $y$  as  $\text{Impact}_{\text{Dir}}(x,y,d)$ . We assume  $d$  to be a numerical objective measure of agent  $y$ ’s performance, with negative  $d$  indicating poor performance. Trust declines when errors occur [33]. However, the size of the error does not seem to be proportional to the decline in trust, with small errors having a larger than expected effect [20, 25]. We assume that the effect on trust of good performance is similar. We define

$$\text{Impact}_{\text{Dir}}(x,y,d) = -\lambda^- + \mu^- \times d, \text{ if } d < 0; \quad \lambda^+ + \mu^+ \times d, \text{ if } d > 0; \quad 0, \text{ if } d = 0$$

with parameters  $\lambda^-/\lambda^+$  (both  $>0$ ) modeling the decline/increase in trust occurring with unsatisfactory/satisfactory performance *independent* of how poor/good the performance was, and  $\mu^-/\mu^+$  (both  $\geq 0$ ) modeling how much the extent of the bad/good performance contributes.

The existing trust of  $x$  in  $y$  (called  $t_{xy}$ ) may affect the impact of  $y$ 's performance on the trust placed in it. It has been hypothesized that the stronger the emotional bond between those in a trust relationship, the less likely contrary behavioral evidence will weaken the relationship [27]. So, we assume poor performance will have less impact if the existing trust is higher. Similarly to the treatment of assimilation in [28], we define

$$\text{Impact}_{\text{Dir}}(x,y,d,t_{xy}) = \text{Impact}_{\text{Dir}}(x,y,d) + (t_{xy} - \text{Impact}_{\text{Dir}}(x,y,d)) \times \varepsilon, \quad \text{with } 0 \leq \varepsilon \leq 1$$

Parameter  $\varepsilon$  models the extent to which  $x$ 's existing trust in  $y$  influences  $x$ 's judgement of  $y$ 's performance: with  $\varepsilon=0$  there is no such influence, with  $\varepsilon=1$   $y$ 's performance has no impact at all.

## 2.2 Reputation. "You have heard that I act well"

The impact of a trust report  $t$  by agent  $z$  to agent  $x$  about agent  $y$  clearly needs to depend on the difference between  $t$  and  $t_{xy}$  ( $x$ 's existing trust in  $y$ ). After all, there should be no impact if this difference is zero. Additionally, the trustworthiness of reporting agent  $z$  ( $t_{xz}$ ) may influence the impact of the trust report<sup>1</sup>. This effect is described in detail in [18], who propose to propagate trust along chains<sup>2</sup>. Table 1 shows seven options for defining  $\text{Impact}_{\text{Rep}}(x,y,z,t,t_{xy},t_{xz})$  which differ in the way  $t_{xz}$  is used. In the first three options (1A-1C),  $t_{xz}$  needs to reach a threshold in order for there to be any impact, but once that threshold is reached, the size of the impact is independent of  $t_{xz}$ . In the latter options, the size of the impact also depends on  $t_{xz}$  (normalized by the maximally obtainable trust  $t_{\text{max}}$ ). Parameter  $\kappa$  is a trustworthiness threshold, e.g. in Option 1A, if  $z$  is trusted more than  $\kappa$  then  $z$ 's report is taken into account, otherwise it is ignored. Parameter  $\varphi$  ( $0 \leq \varphi \leq 1$ ) models an agent's propensity to contagion by others' beliefs: with  $\varphi=0$  trust reports have no impact at all, with  $\varphi=1$  they override the existing trust. We will explore which option is better in Experiment 1.

**Table 1.** Seven options for defining  $\text{Impact}_{\text{Rep}}(x,y,z,t,t_{xy},t_{xz})$

1	A	$\varphi \times (t - t_{xy})$	if $t_{xz} > \kappa$	0 otherwise
	B		if $t_{xz} > t_{xy}$	
	C		if $t_{xz} > \kappa$ and $t_{xz} > t_{xy}$	
2	A	$\varphi \times (t - t_{xy}) \times (t_{xz} / t_{\text{max}})$	if $t_{xz} > \kappa$	
	B		if $t_{xz} > t_{xy}$	
	C		if $t_{xz} > \kappa$ and $t_{xz} > t_{xy}$	
	D		Always	

Often an agent will have multiple experiences with another agent, some of which may be direct and some indirect. This may lead to problems when the existing trust differs a lot from the trust reported by another agent. For instance, suppose agent  $z$  reports distrust in  $y$ , while  $x$  trusts  $y$ , then it is possible that  $x$  may start liking and trusting  $z$  less (instead of or in addition to decreasing trust in  $y$ ). This corresponds to ideas from Congruity theory [36] in the area of persuasion. We will denote the impact on an agent  $x$ 's trust in  $z$  of the trust  $t$  in a third agent  $y$  reported by  $z$ , given existing trust  $t_{xz}$  and  $t_{xy}$  as  $\text{Impact}_{\text{RepGiven}}(x,z,y,t,t_{xz},t_{xy})$ . We define

<sup>1</sup> In fact, two types of trust are involved in reputation: the trust placed in  $z$  is trust as a recommender of other agents (e.g. recommend a plumber), while the trust in  $y$  is trust as performer of some task (e.g. plumbing). Here, we do not explicitly mention the domain of trust, but the model can be easily modified to do this.

<sup>2</sup> In [18], the situation is even more complicated, as they incorporate  $x$  getting a recommendation of  $w$  from  $y$  who in its turn got the recommendation from  $z$  etc.

$$\text{Impact}_{\text{RepGiven}}(x,z,y, t_{xz}, t_{xy}) = -\rho - \tau \times |t - t_{xy}|, \quad \text{if } |t - t_{xy}| > \xi.$$

with  $\rho$  modeling the decline in trust occurring when there is a larger than  $\xi$  discrepancy between  $t$  and  $t_{xy}$  independent of how large this discrepancy really is, and  $\tau$  modeling how much the extent of the discrepancy contributes. We have not yet incorporated  $t_{xz}$  into this formula. We will investigate this further in Experiment 1 below.

The situation may well be more complicated than modelled. The question arises whether agents should remember the trust reports. For instance, suppose  $x$  started trusting  $y$  because it was told that  $y$  was trustworthy by  $z$  whom it trusted at the time. Suppose that later experience shows  $z$  is not trustworthy. Should  $x$  now remember that its trust of  $y$  was based on its trust of  $z$  and revoke its trust in  $y$ ? Similarly, suppose that  $x$  was told  $w$  was trustworthy by  $v$ , but later experience shows  $w$  is not trustworthy. Should  $x$  now reduce its trust in  $v$ ? For simplicity reasons, in this paper we have decided not to use a memory of trust reports. We intend to investigate this issue further in future.

### 3 Experiment 1: Reputation

#### 3.1 Experimental Design

Twenty-six subjects participated in the experiment (16 male, 10 female, average age 32.5, stdev 9.6). All subjects were associated with the computing science department. Each subject was given two scenarios: in each, they were told their trust in two people (say  $Z$  and  $Y$ , names differed per scenario), using a scale from 1 (very untrustworthy) to 10 (very trustworthy)<sup>3</sup>. Next, they were told how much  $Z$  trusts  $Y$ . Subjects rated how trustworthy they now regarded  $Z$  and  $Y$ . The two scenarios were the same except that one scenario (H) would have a higher initial trust rating of  $Z$  than the other (L). The order of the scenarios was randomized.

Subjects were split into groups for three experimental conditions (G1, G2, G3), using different initial trust ratings for  $Z$  and  $Y$ , and a different rating for  $Z$ 's reported trust in  $Y$ . See Table 2 for details. We used a combination of a within- and a between-subjects design. The two scenarios within each group will be used to investigate whether a higher trust in  $Z$  leads to a higher impact of  $Z$ 's trust in  $Y$ . Group G3 differed from G1 by having a very low rating for  $Z$ 's reported trust in  $Y$ . The difference between these two groups will be used to explore the impact on both trust in  $Y$  and  $Z$  of a large difference between the reported trust and the existing trust. Group G2 differed from G1 by having a low rating for the existing trust in  $Z$  in one scenario (4, so untrustworthy) compared to a high rating in a scenario in G1 (10, so very trustworthy). This will be used to explore whether a threshold  $\kappa$  is used, and how low trust in  $Z$  affects the impact of the trust report.

**Table 2.** Trust ratings used in each experimental condition and in each scenario, and results

	Scen	Trust ratings Provided			Predictions about change in trust by modelling options					Resulting Trust: Mean (StDev)		
		Z	Y	Z's in Y	1A	1BC	2A	2BC	2D	Z	Y	
G1	L	7	7	5	-2 $\phi$	0	-1.4 $\phi$	0	-1.4 $\phi$	6.6 (.7)	6.4 (.7)	
	H	10	7	5	-2 $\phi$					9.4 (1.1)	5.6 (.5)	
G2	L	4	7	5	0					-0.8 $\phi$	3.7 (.5)	6.8 (.4)
	H	7	7	5	-2 $\phi$	0	-1.4 $\phi$	0	-1.4 $\phi$	6.8 (.4)	6.6 (.5)	
G3	L	7	7	1	-6 $\phi$	0	-4.2 $\phi$	0	-4.2 $\phi$	6.2 (1.1)	5.7 (1.5)	
	H	10	7	1	-6 $\phi$					9.7 (.5)	4.8 (2.2)	

<sup>3</sup> Telling people their trust on a 1-10 scale may seem unnatural, but is a good simulation of how reputations are often given in an on-line system where users do not know each other. Of course, telling people how much they trust somebody does not necessarily create the same situation as when this degree of trust has arisen through real experiences. However, it seemed a good way to ensure everybody started with the appropriate degree of trust. We will try to do more direct experiments in future.

For each scenario, we calculated the predictions of the options given in Table 1. We had to make some assumptions to do this, as these options contain parameters. For trustworthiness threshold  $\kappa$ , we assumed  $4 < \kappa < 7^4$ . Parameter  $t_{\max}=10$ , as this is the maximal trust available on the scale given to subjects. The trust ratings provided, together with this value of  $\kappa$ , resulted in no differences between the predictions of options 1B and 1C, and no differences between 2B and 2C.

### 3.2 Results and Discussion

Table 2 shows the results of the experiment. Over all subjects, trust in Y decreased in both types of scenario (one sample t-tests,  $p < .005$ ). Comparing this to the predictions by the modelling options in Table 2, this clearly conflicts with 1BC and 2BC. So, it seems that the existing trust in Z being smaller than or equal to that in Y is not resulting in subjects ignoring the trust report. However, there are exceptions: four subjects did indeed mention an effect of the equal trust in Z and Y. For two this resulted in no impact of the trust report: e.g., “Z’s opinion will not affect mine since I trust Z and Y at the same level”. The other two reduced their trust in both Y and Z: e.g., “One of the two is less trustworthy than I thought but I do not know who, so I’m hedging my bets”.

There is also qualitative evidence that subjects used a trust threshold (as advocated in options A). Four subjects mentioned distrust of Z (when it was 4) as a reason for not changing trust in Y: e.g., “I don’t trust Z enough for him to change my opinion”. However, another subject pointed out that trust in Y might suffer slightly nevertheless as “maybe there is [...] some side of him I need to be aware of”. A value of the threshold  $\kappa$  between 4 and 7 seems to correspond with the behaviour and comments of many subjects. However, one subject seemed to put the threshold higher than 7: “I didn’t totally trust Z anyway, so would use my own judgement about Y”. Another three subjects (in G2) indicated that their trust would not be affected by the opinions of others, which may also be caused by using a threshold higher than 7 (the highest score for Z’s trust in G2).

Over all subjects, trust in Y decreased more in the scenario with the higher trust in Z (paired t-test,  $p < .02$ ). Four subjects mentioned their high trust in Z (when it was 10) as a reason for changing trust in Y: e.g., “because I have such a high trust in Z I agree with him in his views”. The degree of trust in Z clearly affects the impact of the trust report. So, the results correspond better with the predictions by options 2A,D than those by option 1A.

Comparing groups G1 and G3 (which only differed in reported trust), the trend is for a higher decrease in trust in Y when Z reports a lower trust (means of 1.4 versus 2.2 for the high trust scenario, .6 versus 1.3 for the low trust scenario). However, the groups are too small for statistical significance. Nevertheless, the data seems to suggest that the decrease is not simply proportional to the trust reported by Z. If it were, we would have expected the decrease to be three times as large in G3 as in G1, which it clearly is not. So, while options 2A and 2D seem best from the ones we proposed above, we change our modelling of  $\text{Impact}_{\text{Rep}}(x,y,z,t,t_{xy},t_{xz})$  within those conditions to:

$$(-\omega + \varphi \times (t - t_{xy})) \times (t_{xz} / t_{\max}), \text{ if } t < t_{xy}; \quad (\omega + \varphi \times (t - t_{xy})) \times (t_{xz} / t_{\max}), \text{ if } t > t_{xy}; \quad 0 \text{ otherwise, with } \omega > 0$$

Over all subjects, trust in Z decreased in both types of scenario (one sample t-tests,  $p < .01$ ). So, the discrepancy between Z’s reported trust in Y and the existing trust in Y leads indeed to a decrease in trust in Z. This is substantiated by the qualitative data. Four subjects (three in G1 and one in G2) defended lowering their trust in Z with reasons like “If Z does not share my trust in Y then I trust Z less”, “Because [...] Z does not share my view about Y”, “Z maybe correct but it is strange he tells me this; it might be some animosity between him and Y”, and “If Z denigrates Y, I would be unsure if I could trust Z myself”. Comparing groups G1 and G3 (which only differed in the size of the discrepancy between the trust report and Y’s existing trust), we did not find any difference in the decrease in trust in Z (the trends for the Low and High trust scenarios are even in directions opposite to each other). So, a discrepancy leads to a decrease in trust, but the size of the

<sup>4</sup> A different scale is used than in the model, where a negative number indicated negative trust. A value of  $4 \leq \kappa \leq 7$  corresponds to  $-1.5 \leq \kappa \leq 2$  on the model’s scale. The scale difference does not matter for our results.

discrepancy does not seem to greatly matter. Therefore, we simplify our modelling of  $\text{Impact}_{\text{RepGiven}}(x,z,y, t_{xz}, t_{xy})$  to

$$\text{Impact}_{\text{RepGiven}}(x,z,y, t_{xz}, t_{xy}) = -\rho, \quad \text{if } |t - t_{xy}| > \xi$$

As we found that a discrepancy of 2 already leads to a decrease in trust, it seems that  $\xi < 2$ .

Over all subjects in all groups, we found no significant difference between the decrease in trust in Z between the two scenarios (mean decrease in trust is .4 for the higher trust scenarios, and a very similar .5 for the lower ones). So, there does not seem to be a need of using  $t_{xz}$  in the calculation of  $\text{Impact}_{\text{RepGiven}}(x,z,y, t_{xz}, t_{xy})$ .

Four subjects (one in G2 and three in G3) said that they needed evidence to back up the trust reports. As one of them put it: “Unless Z tells me a specific story about something Y did, I’m unlikely to revise my trust in Y”. Note that this subject gave this reason when Z had a trust score of 7, not for the other scenario when the trust score was 10. Also, in G3 the trust report is 1, very far from the original trust, so this might require more evidence.

## 4 Modelling stereotypes, empathy, overall trust and user characteristics

### 4.1 Stereotypes. “You presume that I act well”

With neither direct nor indirect evidence at its disposal, an agent may still assume that another agent will act well, based on ‘intuitions’. This is similar to humans drawing conclusions about people on first sight, like some managers who decide within seconds whether to reject an applicant having come to interview. A number of factors influence this feeling:

- *Appearance*. People studying web credibility for various domains (like finance, health, and travel websites) have found a remarkably high influence on credibility of user interface issues (like usability, cool colour tones, balanced layout, or adding a formal author photograph) [9,12,21]. In a sense, one could regard the interface as the cloths of a computer programme. Fogg et al [12] hypothesize that people assume more effort has gone into making a sleek website, and that therefore the author has more to lose by acting badly.
- *Category-based trust*. Humans may trust people belonging to a certain social or organisational grouping more, without necessarily being aware of this bias [24]. For instance, they may place more trust in females (as found by [34] who hypothesised that this may be due to role differences) or doctors. Category-based trust seems similar to “presumed credibility” mentioned by [11] as one of four types of credibility, and illustrated with a negative view of the trustworthiness of car salesmen.
- *Expertise*. Outwardly signs of expertise may increase trust. For instance, it has been found that giving a computer the label of “specialist” made it more credible (as reported in [11]). This is related to so-called role-based trust [24]: if somebody has a certain role in an organisation (i.e. a job title), then people assume they know how to do that job and therefore may trust them more.

We assume that this stereotypical effect on trust happens initially, when the agents first meet. So, it suffices to model it in the initial trust of an agent in another agent, before any (direct or indirect) experiences. We denote the trust produced by  $y$ ’s appearance as  $T_{\text{Appearance}}(y)$ , by  $y$ ’s category membership as  $T_{\text{Category}}(y)$ , by  $y$ ’s expertise as  $T_{\text{Expertise}}(y)$ , and the initial trust of  $x$  in  $y$  due to stereotyping as  $\text{InitialTrust}(x,y)$ . We define

$$\text{InitialTrust}(x,y) = \psi_1 T_{\text{Appearance}}(y) + \psi_2 T_{\text{Category}}(y) + \psi_3 T_{\text{Expertise}}(y), \quad \text{with } \sum \psi_i = 1 \text{ and } \psi_i \geq 0$$

with  $\psi_i$  modelling the relative influence of the factors.  $T_{\text{Appearance}}$ ,  $T_{\text{Category}}$ , and  $T_{\text{Expertise}}$  are clearly domain and user-interface dependent. Defining them further is outside the scope of this paper. We are currently investigating the influence of appearance and expertise on trust in the context of a persuasive health-advice system [34]. We expect that the system designer will

optimize its user interface to have as high an initial trust value as possible. Our model requires this value, as it affects the impact of a user's subsequent experiences. So, some of its factors may need to be determined as part of the design process.

#### 4.2 Empathy “We understand each other”

Another factor that influences trust is empathy: “one's ability to recognize, perceive and directly experientially feel the emotion of another” (Wikipedia). Empathy can partly be seen as having a stereotypical influence: for instance, empathy based on some types of similarity (like gender, ethnicity and age) may be visible at first sight, so may impact trust even before the trustor has experienced the trustee's behaviour. However, it can also evolve over time. Empathy does not only exist between humans: it is possible to achieve empathy in a computer environment [37], and it has been shown that making artificial support agents empathetic helps [23, 3]. Empathy depends on:

- *Similarity between the trustor and the trustee.* Empathy is strongest between people who identify similarities with others or who share experiences [16]. People with the same gender, same occupation, who are close in age, and use similar expressions are more likely to detect others' feelings accurately [16]. If other people are like ourselves we infer they have the same level of trustworthiness [31]. Trust can be fostered by the development of shared common experiences and a group identity [6].
- *Accurate prediction by the trustee of the trustor's feelings and supportive response.* Both empathic accuracy [16] and supportive response (responding compassionately to another person's distress, [4]) had significant influence on online interpersonal trust [10], but inferring alone was not sufficient, there was a need to be supportive as well.
- *Trustor liking the trustee.* Likability is mentioned by [8] as a factor for predictability and intentionality. Buyers are more confident of their predictions about someone they like [46] and attribute favorable motives to them [43]. Liking has been found to be highly correlated with interpersonal trust [10]. It has also been argued that trust breeds trust; the trustee showing trust in the trustor making the trustor reciprocate that trust [30]. This may be linked to likeability, with likeability increasing with trust given.

We incorporate the initial empathy due to similarity into our definition of InitialTrust:

$$\text{InitialTrust}(x,y) = \psi_1 T_{\text{Appearance}}(y) + \psi_2 T_{\text{Category}}(y) + \psi_3 T_{\text{Expertise}}(y) + \psi_4 \text{Similarity}(x,y)$$

We denote the impact on an agent  $x$ 's trust in  $y$  of an empathic experience  $f$  given existing trust  $t_{xy}$  of  $x$  in  $y$  as  $\text{Impact}_{\text{Emp}}(x,y,f,t_{xy})$ . We assume  $f$  to be a number, with a negative number denoting a negative empathic experience. A negative empathic experience may be, for instance: a bad reaction of  $y$  to  $x$ 's emotions, behaviour by  $y$  that shows dissimilarity to  $x$ , or behaviour that makes  $x$  like  $y$  less (e.g. rudeness). Mapping such behaviour onto an objective number is a complicated problem, which is outside the scope of this paper. Obviously, it depends on the interface: the way the agents interact with each other, and how much emotion they can portray and recognize.

We expect that the effect on trust of empathic experiences is similar to that of direct experiences and trust reports, with a small empathic experience having a higher than proportional impact. Also, we assume that assimilation effects happen, just as in the case of direct experiences. Therefore, we treat the impact of empathic experiences the same as direct experiences, and define

$$\text{Impact}_{\text{Emp}}(x,y,f,t_{xy}) = \text{Impact}_{\text{Dir}}(x,y,f,t_{xy})$$

#### 4.3 Modelling overall trust

We denote the trust of  $x$  in  $y$  after a sequence of experiences  $exps$  as  $\text{Trust}(x,y,exps)$ . We assume that trust behaves like affective states: so, as in [28], it decreases over time, is averaged over experiences, and the impact of a new experience depends on the existing trust. We define:

$$\text{Trust}(x,y,<>) = \text{InitialTrust}(x,y)$$

$$\text{Trust}(x,y,exps + <e>) = (\delta * \text{Trust}(x,y,exps) + \text{Impact}(x,y,e,exps,\delta)) / (\delta + 1)$$

$\text{Impact}(x,y,e,exps,\delta)$  denotes the impact of an experience  $e$  on an agent  $x$ 's trust in  $y$  given a previous sequence of experiences  $exps$  and a value  $\delta$  (the latter two can be used to calculate  $x$ 's existing trust in other agents). We define  $\text{Impact}(x,y,e,exps,\delta)$  as

$$\begin{array}{ll} \text{Impact}_{\text{Dir}}(x,y,d,t_{xy}) & \text{if } e \text{ is a direct experience from } x \text{ of } y \text{ with value } d \\ \text{Impact}_{\text{Rep}}(x,y,z,t,t_{xy},t_{xz}) & \text{if } e \text{ is a trust report from } z \text{ to } x \text{ about } y \text{ with value } t \\ \text{Impact}_{\text{RepGiven}}(x,y,z,t,t_{xy},t_{xz}) & \text{if } e \text{ is a trust report from } y \text{ to } x \text{ about } z \text{ with value } t \\ \text{Impact}_{\text{Emp}}(x,y,f,t_{xy}), & \text{if } e \text{ is an empathetic experience from } x \text{ of } y \text{ with value } f \end{array}$$

With  $t_{xy} = \delta * \text{Trust}(x,y,exps)$ , denoting the existing trust of  $x$  in  $y$ .

#### 4.4 Incorporating user characteristics “You trust in general in this area”

It is likely that most of the parameters in our definitions are user dependent. For instance,  $\delta$  and  $\epsilon$  are likely to be user dependent, following the results in [28], and there was some evidence of  $\kappa$  differing between users in Experiment 1. As empathy depends heavily on similarity, this is also user dependent. In addition, these user characteristics influence trust (as also modelled in [5]):

- *General propensity to trust* (GP). People differ in their general propensity to trust others [7,49, 45], and culture can influence this [50,13]. A number of scales for measuring this propensity have been developed [42,50]. However, [10] did not find a correlation between this propensity and online interpersonal trust, and even found the opposite happening in one scenario. It is also unclear whether this propensity extends to computers. We will investigate this in Experiment 2.
- *User expertise* (UE). Several studies have shown that users trust systems less if they have more domain knowledge [20,26]. The opposite has also been shown: if users have less expertise than they trust the system more [48].
- *User need* (UN). If users have a higher need than they trust more (as reported in [11]). Unfamiliar situations and tasks make the need greater. Trying and failing to solve a problem on their own has been shown to lead to user trust in a computer [47, 48].

We propose to combine these characteristics into an agent  $x$ 's general propensity to trust in the given circumstances, denoted by  $\text{GPTC}(x)$ . We define:

$$\text{GPTC}(x) = \gamma_1 \text{GP}(x) - \gamma_2 \text{UE}(x) + \gamma_3 \text{UN}(x), \quad \text{with } \sum \gamma_i = 1 \text{ and } \gamma_i \geq 0$$

We are not certain whether the minus sign before  $\gamma_2 \text{UE}(x)$  is correct, as the literature reports conflicting evidence about the impact of user expertise. We will investigate this in Experiment 2.  $\text{GPTC}(x)$  will impact the initial trust, so we change our definition  $\text{InitialTrust}(x,y)$  to:

$$\psi_1 \text{TAppearance}(y) + \psi_2 \text{TCategory}(y) + \psi_3 \text{TExpertise}(y) + \psi_4 \text{Similarity}(x,y) + \psi_5 \text{GPTC}(x)$$

In addition, it may affect  $\text{Impact}_{\text{Rep}}$ : we will not describe this here in detail, but it will be used to make the parameters in the modelling of  $\text{Impact}_{\text{Rep}}$  user dependent.

## 5. Experiment 2: General propensity to trust and domain expertise

### 5.1 Experimental Design

Twenty-three subjects participated in the experiment (74% male, average age 33, stdev=9.4, all associated with the computing science department). First, subjects' general propensity to trust others was determined using the test in [50] which consists of eight statements to be rated on a 5-

point Likert scale. Secondly, subjects rated their expertise in computing and in five tasks (choosing pop music to listen to, classical music to listen to, wine to buy, a house to buy, and movies to watch) using a 7-point Likert scale (from “little expertise” to “a lot of expertise”). Thirdly, subjects rated how much they would trust a recommender system for each of these tasks using a 7-point Likert scale (from “very low trust” to “very high trust”). Finally, subjects rated for recommender systems in general, how much their trust would be influenced by their own experiences with the system (Exp), its reputation (Rep), and the extent to which they understood how it works (Tr). They also indicated whether this would differ for the systems mentioned.

## 5.2 Results and Discussion

Table 3 shows the results. Using Pearson, we found no significant correlation between subjects’ GP and subjects’ trust in the various recommender systems (correlations were all positive and small, around .2). We did, however, find significant correlations between subjects’ trust in the various recommender systems, for instance trust in recommendations of pop was correlated to trust in recommendations of classical music ( $r=.7$ ,  $p<.001$ ), wine ( $r=.8$ ,  $p<.001$ ), houses ( $r=.5$ ,  $p<.05$ ), and movies ( $r=.9$ ,  $p<.001$ ). So, though we did not find an impact of GP, there may well be an impact of general propensity to trust recommender systems<sup>5</sup>. So, in the model we replace GP with a general propensity to trust the systems in question (GPS).

**Table 3.** Mean (stdev) for the results of Experiment 2

GP	Expertise					Trust in recommender system					Influenced by		
	Pop	Clas	Win	Hou	Mov	Pop	Clas	Win	Hou	Mov	Exp	Rep	Tr
23.1	4.0	3.2	2.9	2.7	4.4	4.0	3.6	4.1	3.6	4.3	5.6	4.5	5.4
(4.6)	(2.0)	(1.9)	(1.6)	(1.8)	(1.3)	(1.1)	(1.5)	(1.3)	(1.5)	(1.1)	(1.4)	(1.2)	(1.2)

We found **no** significant correlations between subjects’ expertise in a topic and their trust in a recommender system on that topic (correlations were all positive and small). Some of the literature discussed earlier indicated that more expertise would lead to lower trust. Indeed, one subject mentioned “I would sooner trust a system on topics about which I have no strong opinion/background knowledge, such as wine”. However, in the overall results, we clearly did not find this, and the trends are even in the opposite direction. Therefore, we have decided to remove the effect of user expertise on the general propensity to trust in the given circumstances.

We had assumed that subjects’ trust in a system recommending houses would be lower, as houses have a high financial and emotional value. Indeed, four subjects explicitly mentioned attributing more trust to a system recommending smaller items. However, the overall results did not confirm this (trust in recommenders of houses and classical music was quite comparable). Maybe this is because the risk a user takes is low: they would go and see a recommended house anyway before buying it. Given the results, we do not incorporate item value into the model.

As shown in Table 3, subjects expected that their trust would indeed be influenced by direct experiences, reputation, and system transparency. They thought direct experiences and transparency would be more important than reputation (pairwise t-tests,  $p<.001$  and  $p<.05$  respectively). Five subjects indicated this would differ a lot depending on the domain of the recommender system, and ten thought it would differ a little. One subject mentioned that reputation and transparency would be more important for domains about which they knew little. Another said transparency was more important for high value recommendations (like houses). A third thought music is more a matter of taste than houses, and that therefore reputation would be more useful for a house recommender than a music recommender. The latter is an interesting point, and it is likely we will need to make parameters in the model domain dependent. Our results also mean that transparency needs to be incorporated in the model.

<sup>5</sup> A confounding factor is that there may be differences between subjects in using rating scales.

## 6 Conclusions

Despite the vast literature on trust across disciplines, there is a lack of *simple computational* models of trust that incorporate the *range of factors* affecting trust. Some computing researchers have presented conceptual models, merely showing the relationship between trust factors (e.g. [5,9,41]). Others, in the field of multi-agent systems, have proposed computational models, but the overview in [44] shows that they model only some factors of trust: ten of the twelve trust models mentioned model only direct experiences and reputation (with four only modelling one of these). The other two incorporate part of what we called stereotypes. The existing models also seem to model at a far more complicated level (using networks of reputation as in [18]). None of them incorporates all the factors discussed here (e.g. they do not incorporate something like  $\text{Impact}_{\text{RepGiven}}$ ). As trust can be seen as a special case of impression formation, research by social psychologists on predicting people's attitude changes (e.g. information integration theory [1]) is also relevant even though not explicitly addressing trust. We already used ideas from Congruity Theory [36]. We intend to produce a comprehensive comparison with existing approaches.

This paper has presented a first step towards a simple, domain independent, computational model of user trust incorporating the main factors affecting trust from the cross-disciplinary literature. We have opted for quantitative equations, rather than trying to model cognitive processing. Clearly more work is needed in this area, and various aspects of the model need to be investigated (and validated) in more detail. For instance, Experiment 1 only looked at negative trust reports, so needs extending to positive ones. To turn the equations into algorithms, parameter values need to be determined, using experiments and user interaction. We would also like to show how this work can be applied to various domains, such as a recommender system domain.

## References

1. Anderson, N.H. (1971). Integration theory and attitude change. *Psychological Review*, 78, 171-206.
2. Ba, S. and Pavlou, P.A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behaviour. *MIS Quarterly*, 26, 243-268.
3. Baylor, A. L., Rosenberg-Kima, R. B., and Plant, E. A. (2006). Interface agents as social models: The impact of appearance on females' attitude toward engineering. *CHI*.
4. Coke, S., Batson, D., and McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *J. of Personality and Social Psychology*, 36, 752-766.
5. Corritore, C.L., Kracher, B., and Wiedenback, S. (2003). On-line trust: concepts, evolving themes, a model. *Int. J. of Human-Computer Studies*, 58, 737-758.
6. Dawes, R.M. and Thaler, R.H. (1988). Cooperation. *J. of Economic Perspectives*, 2, 187-197.
7. Deutsch, M. (1960). Trust, trustworthiness, and the F-scale. *J. of Abnormal and Social Psych.*, 61, 138-140.
8. Doney, P.M. and Cannon, J.P. (1997). An examination of trust in buyer-seller relationships. *J. of Marketing*, 61, 35-51.
9. Egger, F.N. (2000). "Trust me, I'm an online vendor": Towards a model of trust for e-commerce system design. *CHI*, 101-102.
10. Feng, J., Lazar, J., and Preece, J. (2004). Empathy and online interpersonal trust: A fragile relationship. *Behavior and Information Technology*, 23, 97-106.
11. Fogg, B.J. and Tseng, H. (1999). The elements of computer credibility. *CHI*, 80-87.
12. Fogg, B.J., Soohoo, C., Danielson, D., Marable, L., Stanford, J., Tauber, E. (2003). How do users evaluate the credibility of websites?: A study with over 2500 participants. *Designing for user experiences*. 1-15.
13. Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity*. Free Press: New York.
14. Grabner-Kraeuter, S., and Kaluscha, E.A. (2003). Empirical research in on-line trust: a review and critical assessment. *Int. J. of Human Computer Studies*, 58, 783-812.
15. Hupcey, J.E., Penrod, J., Morse, J.M., and Mitcham, C. (2001). An exploration and advancement of the concept of trust. *J. of Advanced Nursing*, 36, 282-293.
16. Ickes, W. (1993). Empathic accuracy. *J. of Personality*, 61, 587-610.
17. James, H. S. (2002). The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness. *J. of Economic Behavior and Organization*, 47, 291-307.

18. Jøsang, A., Gray, E., Kinatader, M. (2006). Simplification and analysis of transitive trust networks, *Web Intelligence and Agent Systems*, 4, 139-161.
19. Jøsang, A., Ismail, R., and Boyd, C. (in press) A survey of trust and reputation systems for online service provision. *Decision Support Systems*.
20. Kantowitz, B.H., Hanowski, R.J. and Kantowitz, S.C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors*, 39, 164-176.
21. Kim, J. and Moon, J. (1998). Designing towards emotional usability in customer interfaces: Trustworthiness of cyber-banking system interfaces. *Interacting with computers*, 10, 1-29.
22. Koehn, D. (1996). Should we trust in trust? *American Business Law Journal*, 34, 183-203.
23. Klein, J., Moon, Y., and Picard, R.W. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14, 119-140.
24. Kramer, R.M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, 50, 569-598.
25. Lee, J. (1991). The dynamics of trust in a supervisory control simulation. *Human Factors Society Annual meeting*, 1228-1232.
26. Lerch, F., Prietula, M. (1989). How do we trust machine advice? Designing and using human-computer interfaces and knowledge-based systems. *Third Annual conf. on Human-Computer-Interaction*. 411-419.
27. Lewis, J.D. and Weigert, A. (1985). Trust as a social reality. *Social Forces*, 63, 967-985.
28. Masthoff, J., and Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems. *UMUAI*, 16, 281-319.
29. McKnight, D.H., and Chervany, N.L. (1996). The meanings of trust. Technical report MISRC Working paper series 96-04, University of Minnesota, Management Information Systems Research Center.
30. Miller, G.J. (1992). *Managerial dilemmas: The political economy of hierarchies*. New York: CUP.
31. Moore, D., Kurtzberg, T., Thompson, L. (1999). Long and short routes to success in electronically mediated negotiations: Group affiliations and good vibrations. *Org. Behav. Human Decision Processes*, 77, 22-43.
32. Moorman, C., Zaltman, G., Deshpandé, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organizations. *J. of Marketing Research*, 29, 314-328.
33. Muir, B.M. and Moray, N. (1996). Trust in automation: Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39, 429-460.
34. Nguyen, H. and Masthoff, J. (forthcoming). Is it me or is it what I say? Source image and persuasion.
35. Orbell, J., Dawes, R., and Schwartz-Shea, P. (1994). Trust, social categories, and individuals: The case of gender. *Motivation and Emotion*, 18, 109-128.
36. Osgood, C., and Tannenbaum, P. (1955). The principle of congruity in the prediction of attitude change. *Psychology Review*, 62, 42-55.
37. Peiris, D.R., Gregor, P., and Alm, N. (2000). The effects of simulating human conversational style in a computer-based interview. *Interacting with Computers*, 12, 635-650.
38. Pu, P. and Chen, L. (2006). Trust building with explanation interfaces. *Int. Conf. on Intelligent User Interfaces*, Sydney, Australia, ACM Press, 93-100.
39. Ramchurn, S.D., Huynh, D., and Jennings, N.R. (2004). Trust in multi-agent systems. *The Knowledge Engineering Review*, 19, 1-25.
40. Resnick, P., Zeckhauser, R., Friedman, E., Kuwabara, K. (2000). Reputation systems. *Com. ACM*, 43, 45-8.
41. Riegelsberger, J., Sasse, M.A., and McCarthy, J.D. (2005). The mechanics of trust: A framework for research and design. *Int. J. Human-Computer Studies*, 62, 381-422.
42. Rotter, J.B. (1967). A new scale for the measurement of interpersonal trust. *J. of Personality*, 35, 651-665.
43. Rotter, J.B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35, 1-7
44. Sabater, J., Sierra, C. (2005). Review on computational trust and reputation models. *AI Review*, 24, 33-60.
45. Sorrentino, R.M., Holmes, J.G., Hanna, S.E., Sharp, A. (1995). Uncertainty orientation and trust in close relationships: Individual differences in cognitive styles. *J. of Personality and Social Psych.*, 68, 314-327.
46. Swan, J.E. and Nolan, J.J. (1985). Gaining customer trust: A conceptual guide for the salesperson. *J. of Personal Selling and Sales Management*, 5, 39-48.
47. Wearn, Y., Hägglund, S., Löwgren, J., Rankin, I., Sololnicki, T. and Steinmann, A. (1992). Communication knowledge for knowledge communication. *Int. J. of Man-Machine Studies*, 37, 215-239.
48. Waern, Y. and Ramberg, R. (1996). People's perception of human and computer advice. *Computers in Human Behavior*, 12, 17-27.
49. Wrightsman, L.S. (1991). Interpersonal trust and attitudes toward human nature. In J. Robinson et al. (Eds.), *Measures of personality and psychological attitudes*, Academic: San Diego, CA, 373-412.
50. Yamagishi, T. (1988). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, 51, 265-271.