Aiding Human Reliance Decision Making Using Computational Models of Trust

Peter-Paul van Maanen
TNO Human Factors, Soesterberg, The Netherlands &
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
peter-paul.vanmaanen@tno.nl

Tomas Klos
Dutch National Research Institute for
Mathematics and Computer Science (CWI),
Amsterdam, The Netherlands
tomas.klos@cwi.nl

Kees van Dongen
TNO Human Factors, Soesterberg, The Netherlands
kees.vandongen@tno.nl

Abstract

This paper involves a human-agent system in which there is an operator charged with a pattern recognition task, using an automated decision aid. The objective is to make this human-agent system operate as effectively as possible. Effectiveness is gained by an increase of appropriate reliance on the operator and the aid. We studied whether it is possible to contribute to this objective by, apart from the operator, letting the aid as well calibrate trust in order to make reliance decisions. In addition, the aid's calibration of trust in reliance decision making capabilities of both the operator and itself is also expected to contribute, through reliance decision making on a metalevel, which we call metareliance decision making. In this paper we present a formalization of these two approaches: a reliance (RDMM) and metareliance decision making model (MetaRDMM), respectively. A combination of laboratory and simulation experiments shows significant improvements compared to reliance decision making solely done by operators.

1. Introduction

Human-aid cooperation in complex domains, such as aviation, nuclear power, or health care, is becoming increasingly common. The idea of this is that the performance of humans in closer cooperation with decision aids (agents), and vice versa, perform better than humans or decision aids working separately, without taking the other into account. Although this performance benefit is often observed in human-aid teams, cooperation effectiveness of the decision aid is not always fully realized.

In recent work [2, 10] a human-aid team was studied where a human operator, charged with a pattern recognition task, was supported by an automated decision aid. The objective of the task was to make this human-aid team operate as effectively as possible. It turned out that in many occasions the operator made wrong reliance decisions and therefore effectiveness decreased.

Ideally humans rely on their own decisions when these are best and rely on the decision aid's when those are best. But operators cannot be expected to base their reliance decisions on comparisons of true reliabilities of themselves and those of the decision aids. Rather, perceived reliabilities are used which, unfortunately, are usually imperfectly calibrated to true reliabilities, even after practice [2]. It is often found that humans rely either too much or too little on decision aids or themselves [12, 13, 3, 2].

People use relative trust to decide whom to rely on [11]. Trust is defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability [9]. Trust can refer to the advice of another agent or to one's own judgment. Trust, like the perceptions of reliability on which it is based, is a covert or cognitive state [4].

Perceptions of reliability may be prone to systematic and random error. One such error is overtrust: humans may overestimate their own performance or that of the aid. Humans are for instance known to overestimate the number of tasks they can complete in a given period of time [1]. Another error is undertrust: humans may underestimate their own performance or that of the aid. Concerning the perception of the aid's performance, in [14, 2] it was for instance found that the reliability of decision aids is often underestimated. When the direction or magnitude of such errors differ between self and aid, this could lead to inappropri-

0-7695-3028-1/07 $25.00 © 2007 IEEE
DOI 10.1109/WI-IATW.2007.108

ate reliance decisions: Underreliance or overreliance may be the result.

Because the aid is unaffected by cognitive biases, like humans are, the first question raised in this paper is whether it is possible to let the aid make more accurate trust assessments, and therefore reliance decisions, than the operator. In that case, reliance decision making done by the aid is expected to lead to a decrease of over- and underreliance.

Nonetheless, the transparent character of the operator's own motivation for his performance may result in a substantial amount of occasions in which humans make better reliance decisions than aids. In these cases, the suggested reliance decision making completely done by the aid does not result in an optimal performance. The second question therefore raised is whether it is possible to let the aid make even more accurate reliance decisions when based on a prediction if such situations are at hand and then the decision is made to rely on the operator if that is more appropriate. This type of decision making is on a metalevel and therefore is called *meta*reliance decision making. It is expected to result in a further decrease of over- and underreliance.

This paper is composed of several sections addressing the above two questions. First, in Section 2 it is shown how an extension of the task environment from [2] is used as a base for studying the effectiveness of aiding human reliance decision making. Decision aid design and the formalization of the reliance decision making models used by the aid, i.e. a *reliance (RDMM)* and *metareliance decision making model (MetaRDMM)*, respectively, are presented in Section 3. Section 4 describes the method of the experiment and simulation done. The results in terms of model performance by comparison with operator performance from [2] are presented in Section 5. Section 6 ends this paper with some conclusions and suggestions for further research.

2. Task environment

For the experiment described in [2] participants read a story about a software company interested in evaluating the performance of their adaptive software before applying it to more complex tasks on naval ships. Participants were asked to perform a pattern recognition task with advice of a decision aid and were instructed to maximize the number of correct answers by either relying on their own or the decision aid's predictions.

The interface of the task contained 4 rows. Each row consisted of a progress bar, buttons numbered 1, 2, and 3, and a phase description. In Phase 1 the operator had to predict which button to push, based on what they thought the pattern was. In Phase 2 the aid had to do the same. In Phase 3 the operator again had to decide which button to push, this time also taking the prediction of the aid into account, which required the operator also to make a reliance decision. In the final phase feedback was given on what button was correct. Each experiment contained 101 trials, each consisting of these four phases.

In order to support the operator in making reliance decisions the above rows were extended to a total of six (see example interaction in Figure 1), which means two phases were added: In Phase 4 the aid had to make a reliance decision similar as the operator's in Phase 3. In Phase 5 the aid had to decide when to follow the reliance decision of the operator and when its own. These kind of decisions are called *meta*reliance decisions. After this, Phase 6 was the feedback phase.

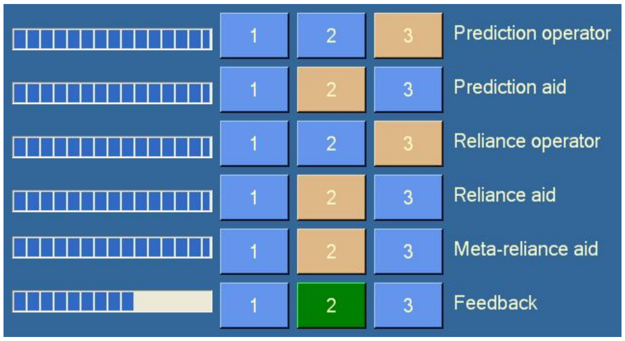

Figure 1. Example operator-aid interaction. The rows represent different phases.

In Figure 1 the following scenario is shown: the operator predicts number 3 (Phase 1), the aid number 2 (Phase 2), then the operator wants to rely on himself (Phase 3), the aid also relies on itself (Phase 4), then the aid decides to *meta*rely on itself again (Phase 5), and finally the feedback shows this was the appropriate decision (Phase 6). Both interpret the outcome and go on to the next trial.

Note that no other support than mentioned above is given to the operator (e.g., no correct answer history is kept for the operator and the operator was not allowed to write things down). Feedback is based on a predefined but then partly randomized sequence of the numbers 1, 2, and 3. The predictions of the aid were also predefined. Each participant got a comparable but different sequence. See Section 4 for more details.

3. Decision aid design

In this section the design of the aid is described in terms of how the aid's decisions in Phases 4 (RDMM) and 5 (MetaRDMM) are made (see end of Section 4 for details on Phase 2), building on [7, 10]. In these phases the aid estimates and compares the task-related (prediction) and reliance decision making capabilities, respectively, of the operator and itself. The idea is to let the aid estimate its trust in the operator's and its own prediction and reliance decision making capabilities each time that feedback is given in Phase 6. As a model, we use a Beta probability density function (pdf) over the different values that the agents' (operator's or aid's) respective capabilities can have. Upon receiving the feedback in Phase 6, the aid uses Bayes' rule to update its estimations [5, 6, 7, see [8] for a generalization to the Dirichlet distribution].

From the perspective of the aid, each agent's behavior can be seen as a sequence of Bernouilli trials, governed by a bias or probability of the outcome 'success' in each trial, called θ_a^x for $x \in \{\texttt{prediction}, \texttt{reliance}\}$ and $a \in \{\texttt{operator}, \texttt{aid}\}$. It is this probability that the aid needs to estimate for the two possible values of both x and a. For $x =$ prediction, this yields two values for RDMM, and for $x =$ reliance, it yields two values for MetaRDMM. In the remainder of this section we drop the sub- and superscripts a and x.

The probability of n successes in N Bernouilli trials ($0 \leq n \leq N$) is given by the Binomial probability mass function

$$p(n|\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}. \tag{1}$$

This also gives the Binomial likelihood of θ, when interpreted as a function of the second argument θ with n held fixed. This likelihood may be used to update the posterior probability $p(\theta|n)$, using Bayes' rule:

$$p(\theta|n) = \frac{p(n|\theta)p(\theta)}{p(n)}. \tag{2}$$

The Beta pdf is a conjugate prior for the Binomial likelihood, which means that if it is used as the prior $p(\theta)$ in Eq. 2, the posterior $p(\theta|n)$ is again a Beta pdf. The Beta pdf is the following:

$$\mathrm{Beta}(\theta|r,s) = \frac{1}{\beta(r,s)} \theta^{r-1}(1-\theta)^{s-1}, \tag{3}$$

for $0 \leq \theta \leq 1$ and $s, r > 0$, where $\beta(r,s)$ is the beta function, and s and r are the number of successes and failures, respectively. [1] The expected value of the Beta distribution is $E(\theta) = \frac{r}{r+s}$.

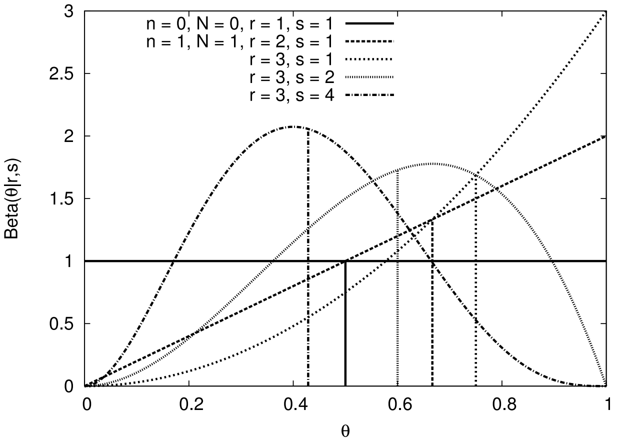

Figure 2. The Beta pdf of θ for different values of r and s.

As explained above, the posterior distribution is still a Beta distribution (disregarding the normalization factor in the denominator of Bayes' rule, since it does not depend on θ):

$$\overbrace{p(\theta|n,N,r,s)}^{\text{posterior}} \propto \overbrace{\left[\theta^n(1-\theta)^{N-n}\right]}^{\text{likelihood (see Eq. 1)}} \overbrace{\left[\theta^{r-1}(1-\theta)^{s-1}\right]}^{\text{prior (see Eq. 3)}}$$

$$\propto \theta^{n+r-1}(1-\theta)^{N-n+s-1},$$

with expected value $E(\theta) = \frac{r+n}{r+s+N}$. In effect, one simply adds the new counts of successes (n) and failures ($N-n$) to the old values of the parameters of the Beta distribution r and s, respectively, and obtains a new distribution with parameters r' and s'.

In the context of trust models, an agent i's trust $\tau_i(j)$ in another agent j's capabilities or intentions, is usually calculated as the expected value of the beta function $\beta(u+1, v+1)$, where u and v are the current counts of positive and negative experiences i has had with j. In the absence of such experiences, the values $r = s = 1$ are typically used for binary outcomes, yielding a uniform prior, and an expected value of 0.5 for the value estimated to govern j's behavior. Updating this uniform prior with positive an negative evidence u and v, respectively, yields $E(\theta) = \frac{u+1}{u+v+2}$. Figure 2 gives the shape of the Beta probability density function of θ given different amounts of evidence, where the expected values of θ are indicated by vertical lines. Because we have 3 possible outcomes in each phase, we initialize the prior as $p = \frac{1}{3}$, by setting $r = 1$ and $s = 2$. Furthermore, we discount old evidence [6], by using a discount factor $0 \leq \lambda \leq 1$, with which old evidence is multiplied before new evidence is added.

For each trial, when the two agents' predictions or reliances differ, (Meta)RDMM selects the prediction (in Phase 4) or reliance (in Phase 5) made by the most highly trusted agent, using four updated trust values $\tau_{\texttt{aid}}^x(a)$ of the aid. In Figure 3 the aid's trust dynamics for an arbitrary operator are shown. For trial 14, for instance, the phase outcomes are similar as in the scenario shown in Figure 1: For this trial, the operator predicted 3, the aid 2, the operator relied on himself, and the correct number was 2 (Phases 1–

[1] The beta function is

$$\beta(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)},$$

where $\Gamma(x) = (x-1)!$ is the Gamma function, where x is a positive integer.

3, 6). Because the operator prediction trust is lower than the aid prediction trust ($\tau^{\texttt{prediction}}_{\texttt{aid}}(\texttt{operator}) < \tau^{\texttt{prediction}}_{\texttt{aid}}(\texttt{aid})$), the aid relied on itself (Phase 4), and because the operator reliance trust was lower than the aid reliance trust ($\tau^{\texttt{reliance}}_{\texttt{aid}}(\texttt{operator}) < \tau^{\texttt{reliance}}_{\texttt{aid}}(\texttt{aid})$), the aid also *meta*relied on itself (Phase 5).

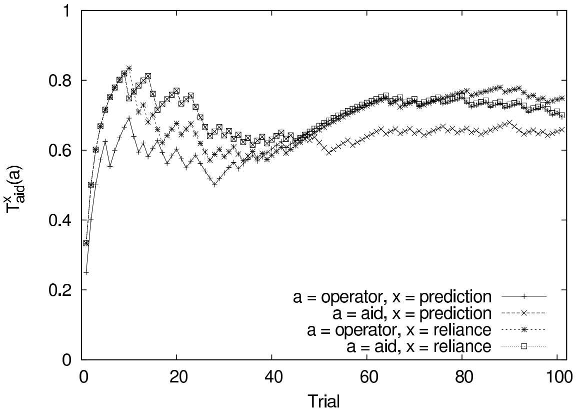

Figure 3. Example of different trust dynamics for an arbitrary operator.

4. Method

Participants The experimental data related to the input of the operator (Phases 1 and 3) were taken from the experiments described in [2]. Forty three Dutch university students (16 female, 18–38 yrs, $M = 23$ yrs) participated in the experiment. Participants were paid €35 for their participation. To control for learning effects, each participant performed the same experiment twice. This means there were a total of 86 experiments, each containing 101 trials. The decision aid was *simulated* offline to be aiding these participants as described in Section 2.

Design Performances for three phases were calculated: the operator's reliance phase (OperatorRDM), the aid's reliance phase (RDMM), and the aid's metareliance phase (MetaRDMM), i.e. Phases 3–5. Only those trials were interesting in which either the operator or (exclusive or) the aid made a correct prediction or reliance decision. This is due to the fact that, in the case of prediction and reliance consensus and in the situation where neither operator nor aid is correct in their prediction or reliance, comparison of aid and operator performance is uninformative.[2] The independent variables for each performance measure were operator and aid prediction accuracy (for Phases 1 and 2), which are described below in more detail.

Operator prediction accuracy was manipulated by varying the difficulty of predicting a predefined sequence of the numbers 1, 2, and 3. The order of the predefined sequence determined the order of the given feedback in Phase 6, which was the only source for the participants to learn the sequence. The predefined sequence was a repeated, but randomized, pattern of length 5. Note here that participants did not know they were subject to identification of a (randomized) repeated sequence. Due to the fact that humans tend to see patterns in noise and because of a convincing story told in the beginning of the experiment, they rather thought it was a sequence dependent on certain complex patterns still to be discovered by them. This has also been confirmed by a post-experimental questionnaire.

Aid prediction accuracy manipulation was based on randomization of the above mentioned randomized predefined sequence. The accuracy of the aid was set on average at 70% ($SD = 3\%$), which is similar to the expected operator prediction accuracy. This was done to make reliance decision making nontrivial for the operator.

5. Results

Based on the experiments it is found that on average, for each participant, in 47.64% ($SD = 6.23\%$) of all trials either the operator ($M = 34.19\%$, $SD = 12.44\%$) or the aid ($M = 65.81\%$, $SD = 12.44\%$) predicted correctly. These last two averages differ substantially from 0 ($N = 48$, $p = 0.00$), which suggests that optimal performance is not reached simply by relying only on the aid or operator. For the mentioned trials, the performances (percentages correct) of OperatorRDM ($M = 58.65\%$, $SD = 9.79\%$) and RDMM ($M = 66.38\%$, $SD = 10.43\%$) are shown in Figure 4 (empty bars). The RDMM results show a significant improvement compared to OperatorRDM ($t = 4.98$, $p = 0.00$).

On average, for each participant, in 22.04% ($SD = 9.88\%$) of all trials either the operator ($M = 40.85\%$, $SD = 18.28\%$) or the aid ($M = 59.15\%$, $SD = 18.28\%$) relied correctly. These last two averages differ substantially from 0 ($N = 22$, $p = 0.00$), which suggests that reliance decision making completely done by the aid does not result in an optimal performance. For the mentioned trials, the performances of OperatorRDM, RDMM, and MetaRDMM ($M = 59.80\%$, $SD = 16.81\%$) are shown in Figure 4 (pattern bars). The MetaRDMM results show a significant improvement compared to OperatorRDM ($t = 7.03$, $p = 0.00$) and an insignificant improvement compared to RDMM ($t = 0.24, p = 0.81$). There is no significant difference between the two experiments per participant. Hence,

[2] Although it appears that in the experiments 0.64% of the trials participants decided not to, or were too late to, rely on prediction consensus, it had no significant influence on the present results.

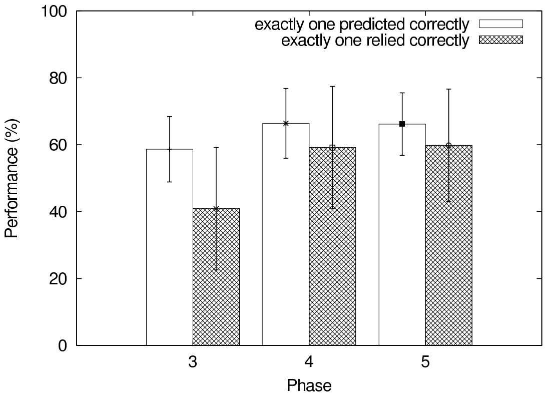

Figure 4. OperatorRDM (Phase 3), RDMM (Phase 4), and MetaRDMM (Phase 5) performances.

there are no measurable learning effects.

6. Conclusion

The general goal of this work is to develop concepts that improve performance of human-aid teams. Improvement is reached by aiding human reliance decision making through the usage of computational models of trust. Our results showed significant results in which decision models RDMM and MetaRDMM outperform human reliance decision making capabilities. The participants may have performed worse than (Meta)RDMM because of limited attentional and memory resources and biases in weighing successes and failures of both themselves and the aid.

As was expected, the results still show a substantial amount of occurrences in which humans make better reliance decisions than aids. MetaRDMM tries to take advantage of this. Although our results show that MetaRDMM also outperforms human participants, a significant improvement compared to RDMM was not found. The first research question raised in this paper can thus be answered with yes, but the answer for the second remains a challenge for further research. Results may differ if the experiment is redone using the extended task described in this paper. One of the positive effects MetaRDMM might imply is a lower human performance degradation, and thus a stronger advantage to RDMM.

It is expected that in real world settings both human reliance decision making and the opportunities for support will be different. Humans, for instance, use additional cues for calibrating trust. Also feedback is often not immediately available, is not always accurate, or complete. The application of the presented concepts and models in real world settings must therefore also be subject to further exploration.

Acknowledgments This research is partly funded by the Royal Netherlands Navy under progr. nr. V206 and by the Dutch government (SENTER) under proj. nr. TSIT2021.

References

[1] R. Buehler, D. Griffin, and M. Ross. Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, 67:366–381, 1994.

[2] K. Van Dongen and P.-P. Van Maanen. Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. In *Proceedings of the Human Factors and Ergonomics Society's 50th Annual Meeting*, San Francisco, USA, 2006.

[3] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe. Misuse and disuse of automated aids. In *Proceedings of the Human Factors Society 43rd Annual Meeting*, pages 339–343, Santa Monica, CA, 1999.

[4] R. Falcone and C. Castelfranchi. Social trust: a cognitive approach. *Trust and deception in virtual societies*, pages 55–90, 2001.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2004.

[6] A. Jøsang and R. Ismail. The Beta reputation system. In *Proc. 15^{th} Bled Electronic Commerce Conference*, Bled, Slovenia, June, 17–19 2002.

[7] T. B. Klos and H. La Poutré. A versatile approach to combining trust values for making binary decisions. In *Trust Management*, volume 3986 of *Lecture Notes in Computer Science*, pages 206–220. Springer, 2006.

[8] K. Krukow. *Towards a Theory of Trust for the Global Ubiquitous Computer*. PhD thesis, University of Aarhus, 2006.

[9] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46:50–80, 2004.

[10] P.-P. Van Maanen and K. Van Dongen. Towards task allocation decision support by means of cognitive modeling of trust. In *Proceedings of the Eighth International Workshop on Trust in Agent Societies (Trust 2005)*, pages 168–77, Utrecht, The Netherlands, July 2005.

[11] N. Moray, T. Inagaki, and M. Itoh. Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, 6:44–58, 2000.

[12] R. Parasuraman and V. A. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39:230–253, 1997.

[13] L. J. Skitka, K. L. Mosier, and M. Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

[14] D. A. Wiegmann, A. Rich, and H. Zhang. Automated diagnostic aids: the effects of aid reliability on user's trust and reliance. *Theoretical Issues in Ergonomics Science*, 2:352–367, 2001.