Closed-Loop Adaptive Decision Support Based on Automated Trust Assessment

Peter-Paul van Maanen[1,2], Tomas Klos[3], and Kees van Dongen[1]

[1] TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
{peter-paul.vanmaanen, kees.vandongen}@tno.nl
[2] Department of Artificial Intelligence, Vrije Universiteit Amsterdam,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[3] Dutch National Research Institute for Mathematics and Computer Science (CWI),
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
tomas.klos@cwi.nl

Abstract. This paper argues that it is important to study issues concerning trust and reliance when developing systems that are intended to augment cognition. Operators often under-rely on the help of a support system that provides advice or that performs certain cognitive tasks autonomously. The decision to rely on support seems to be largely determined by the notion of relative trust. However, this decision to rely on support is not always appropriate, especially when support systems are not perfectly reliable. Because the operator's reliability estimations are typically imperfectly aligned or calibrated with the support system's true capabilities, we propose that the aid makes an estimation of the extent of this calibration (under different circumstances) and intervenes accordingly. This system is intended to improve overall performance of the operator-support system as a whole. The possibilities in terms of application of these ideas are explored and an implementation of this concept in an abstract task environment has been used as a case study.

1 Introduction

One of the main challenges of the Augmented Cognition Community is to explore and identify the limitations of human cognitive capabilities and try to let technology seemlessly adapt to them. This paper focuses on augmenting human cognitive capabilities concerning reliance decision making.

Operators often under-rely on the help of a support system that provides advice or that performs certain cognitive tasks autonomously. The decision to rely on support seems to be largely determined by the notion of relative trust. It is commonly believed that when trust in the support system is higher than trust in own performance, operators tend to rely on the system. However, this decision to rely on help is not always appropriate, especially when support systems are not perfectly reliable. One problem is that the reliability of support systems is often under-estimated, increasing the probability that support is rejected. Because the operator's reliability estimations are typically imperfectly aligned or calibrated with true capabilities, we propose that the aid makes an estimation of the extent of this calibration (under different circumstances) and intervenes accordingly. In other words, we study a system that assesses whether human

decisions to rely on support are made appropriately. This system is intended to improve overall performance of the operator-support system as a whole.

We study a system in which there is an operator charged with making decisions, while being supported by an automated decision support system. As mentioned above, the aim is to make the operator-support system as a whole operate as effectively as possible. This is done by letting the system automatically assess its trust in the operator and in itself, and adapt or adjust aspects of the support based on this trust. This requires models of trust, including a way of updating trust based on interaction data, as well as a means for adapting the type of support.

In this study, trust is defined as the attitude that an agent will help achieve an individual's goals, possibly the agent itself, in a situation characterized by uncertainty and vulnerability [1]. Trust can refer to the advice of another agent or to one's own judgment. Trust, like the feelings and perceptions on which it is based, is a covert or psychological state that can be assessed through subjective ratings. To assess trust, some studies have used scales of trust (e.g., [2]) and some studies have used scales of perceived reliability (e.g., [3]). The latter is used because no operator intervention is needed. We distinguish trust from the decision to depend on advice, the act of relying on advice, and the appropriateness of relying on advice [4, 5].

As a first implementation of this closed-loop adaptive decision support system, the operator-system task described in [6] has been extended.[4] This architecture instantiation leads to an overview of the lessons learned and new insights for further development of adaptive systems based on automated trust assessment. The present paper discusses some key concepts for improving the development of systems that are intended to augment cognition. The focus is on improving reliance on support.

In Section 2 an overview is given of the theoretical background of reliance decision making support systems and its relevance to the Augmented Cognition Community. In Section 3 the conceptual design of a reliance decision making support system is given. In Section 4 an instantiation of this design is described and evaluated. We end with some conclusions and future research.

2 Theoretical Background

The goal of augmented cognition is to extend the performance of human-machine systems via development and usage of computational technologies. Adaptive automation may be used to augment cognition. Adaptive automation refers to a machine capable of dynamic reallocation of task responsibility between human and machine. Reallocation can be triggered by changes in task performance, task demands, or assessments of workload. The goal of adaptive automation is to make human-machine systems more resilient by dynamically engaging humans and machines in cognitive tasks. Engaging humans more in tasks may solve out-of-the-loop performance problems, such as problems with complacency, situation awareness, and skills-degradation. This may be useful in situations of underload. Engaging machines more in tasks may solve performance degradation when the demand for speed or attention exceeds the human ability. This may be useful in situations of overload.

[4] A description and analysis of this system will be published in another paper in preparation.

It should be noted that the potential benefits of adaptive automation turn into risks when the system wrongly concludes that support is or is not needed, or when the timing or kind of support is wrong [7]. For the adaptive system there may be problems with the real-time acquisition of data about the subject's cognition, with determining whether actual or anticipated performance degradations are problematic, and with deciding whether, when, and in what way activities need to be reallocated between human and machine. When the adaptive system is not reliable we create rather than solve problems: unwanted interruptions and automation surprises may disrupt performance and may lead to frustration, distrust, and disuse of the adaptive system [8]. In this paper we focus on computational methods that can be used to adjust the degree in which the machine intervenes.

When machine decisions about task reallocation are not reliable under all conditions the human operator should somehow be involved. One way is to make the reasoning of adaptive automation observable and adjustable for the operator. Understanding the machine's reasoning would enable her to give the system more or less room for intervention. Another and more ambitious way to cope with unreliable adaptive automation is by having a machine adjust its level of support based on a real-time model of trust in human reliance decision making capabilities. In this case it is the machine which adjusts the level of support it provides. The idea is adjusting the level of support to a level that is sufficiently reliable for the user, that problems with frustration, distrust and disuse of the adaptive system are reduced.

A rational decision maker accepts support of an adaptive system when this would increase the probability of goal achievement and reject this support when it would decrease goal achievement. We rely more on support when we believe that it is thought to be highly accurate or when we are not confident about our own performance. People seem to use a notion of relative trust to decide whether to seek or accept support [9–11]. We also rely more on support when the decision of the system to provide support corresponds to our own assessment. The performance of an adaptive support system has to be trusted more than our own performance as well as be appropriately timed. In making a decision to accept support, users are thought to take the reliability of past performance into account. This decision to accept support is not based on a perception of actual reliability, but on how this is perceived and interpreted. Unfortunately, research has shown that trust and perceptions of reliability may be imperfectly calibrated: the reliability of decision support is under-estimated [3, 6]. This could lead to under-reliance on systems that provide adaptive support. In this paper we argue that, because of this human bias to under-rely on support, reliance decision support designs are needed that have the following properties:

Feedback They should provide feedback about the reliability of past human and machine performance. This would allow humans to better calibrate their trust in their own performance and that of the machine, and support them to appropriately adjust the level of autonomy of adaptive support.

Reliance They should generate a machine's decision whom to rely on. Humans could use this recommendation to make a better reliance decision. This decision could also be used by the machine itself to adjust its level of autonomy.

Meta-reliance They should generate a machine's decision whom to rely on concerning reliance decisions. This decision could combine and integrate the best reliance decision making capabilities of both human and machine. This could also be used by the machine itself to adjust its level of autonomy.

In the following sections we show how the above three functions could be realized by a system that automatically assesses trust in real-time.

3 Conceptual Design of Reliance Decision Support

In this section the three properties mentioned above are described in more detail, in terms of three increasingly elaborate conceptual designs of reliance decision support. First we abstract away from possible application domains in order to come to a generic solution. The designs presented in this section are applicable if the following conditions are satisfied:

- The application involves a human-machine cooperative setting concerning a complex task, where it is not trivial to determine whether the machine or the human has better performance. In other words, in order to perform the task at hand, it is important to take both the human's and the machine's opinion into account.
- Both the human operator and the automated aid are able to generate solutions to the problems in the application at hand. In other words, both are in principle able to do the job and both solutions are substitutable, but not necessarily generated in a similar way and of the same quality.
- Some sort of feedback is available in order for both machine and human to be able to estimate their respective performances and generate trust accordingly. In other words, there is enough information for reliance decision making.

In many cases, if for a certain task the above conditions do not hold (e.g., the operator's solution to a problem is not directly comparable to the aid's solution, or no immediate feedback is available), then for important subtasks of the task they generally still hold.

One could say that for all automated support systems the aid supports the operator on a scale from a mere advice being asked by the user, to complete autonomous actions performed and initiated by the aid itself. More specifically, for reliance decision making support, this scale runs from receiving advice about a reliance decision, to the reliance decision being made by the aid itself. In a human-machine cooperative setting, a *reliance decision* is made when either the aid or the operator decides to rely on either self or other. In the designs presented below the terms *human advice* and *machine advice* refer to the decision made for a specific task. The terms *human reliance* and *machine reliance* refer to the reliance decisions made by the human and the machine, respectively, i.e., the advice (task decision) by the agent relied upon. Finally, the term *machine meta-reliance* refers to the decision of the machine whether to rely on the human or the machine with respect to their reliance capabilities.

3.1 Feedback

Agreement or disagreement between human and machine concerning their advice can be used as a cue for the reliability of a decision. In case of agreement it is likely that

(the decision based on) the corresponding advice is correct. In case of disagreement, on the other hand, at least one of the advices is incorrect. To decide which advice to rely on in this case, the operator has to have an accurate perception of her own and the aid's reliability in giving advice. The machine could provide feedback about these reliabilities, for instance by communicating past human and machine advice performance. This would allow humans to better calibrate their trust in their own performance and that of the machine, and support them to adjust the machine's level of autonomy. In Figure 1 the conceptual design of machine feedback is shown.

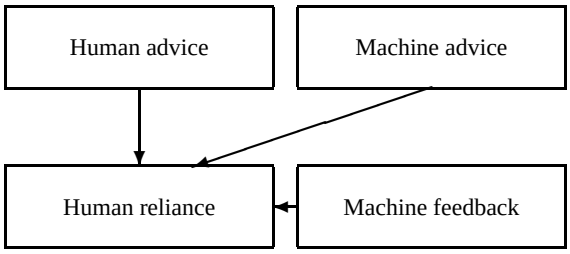

Fig. 1. Both human and machine generate an advice on which the human's reliance decision is based. The machine provides feedback, for instance about the reliability of past human and machine performance. This allows humans to better calibrate their trust.

3.2 Reliance

Unfortunately, by comparing advice, one introduces an extra cognitive task: making a reliance decision. In this particular design the machine augments the cognitive function of reliance decision making, resulting in a decrease of the operator's workload. This can be in the form of a recommendation, or the reliance decision can be made autonomously by the machine, without any intervention by the human operator. The machine or human could adjust the machine's level of autonomy in that sense. Additionally, the human could provide feedback in order to improve the machine's decision. For instance, the human can monitor the machine in its reliance decision making process and possibly veto in certain unacceptable situations. In Figure 2 the conceptual design of such machine reliance is shown.

3.3 Meta-reliance

Since in some situations humans make better reliance decisions, and in others machines do, reliance decision making completely done by the machine does not result in an optimal effect. Therefore, it may be desirable to let the machine decide whom to rely on concerning making reliance decisions. We called this process *meta-reliance decision making* and it combines the best reliance decision making capabilities of both human and machine. If the machine's meta-reliance decision determines that the machine itself should be relied upon, the machine would have a high level of autonomy, and otherwise

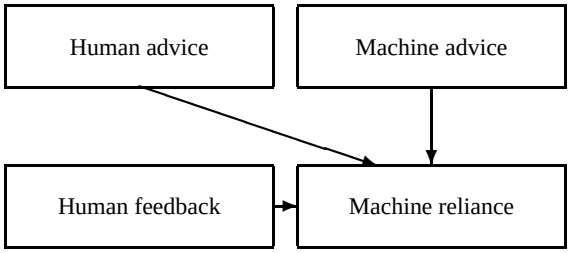

Fig. 2. The machine generates a reliance decision. In this particular design the machine augments the cognitive function of reliance decision making. Both human and machine generate an advice on which the machine's reliance decision is based. It is possible that the human gives additional feedback.

a lower one. Hence the machine is capable of adapting its own autonomy. In Figure 3 the conceptual design of machine meta-reliance is shown.

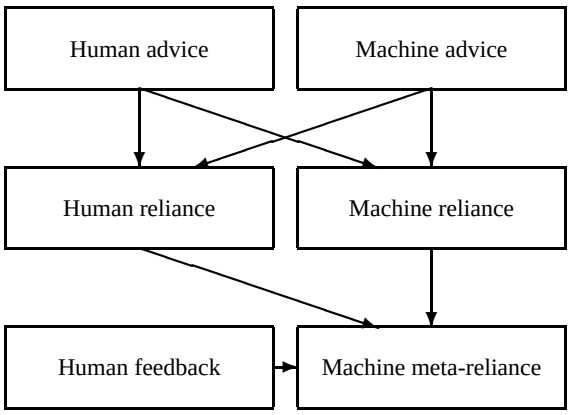

Fig. 3. The machine generates a meta-reliance decision. It combines the best reliance decision making capabilities of both human and machine. Both the human and the machine generate advices and reliance decisions, on the latter of which the machine's meta-reliance decision is based.

4 Implementation and Evaluation

In this section we describe a proof-of-concept for the ideas presented above. In previous work [6], a collaborative operator-aid system was used in laboratory experiments to study human operators' reliance decision making. None of the additions described in Sec. 3 were employed, the setting was essentially that described in Sec. 3.1, without

the aid's feedback. We have now extended the aid's design to provide the reliance and meta-reliance properties, and simulated the extended system's performance, compared to the results from the laboratory experiments. Below, we first describe the original and the extended task, and then the corresponding extensions in the aid's design. Finally, we present the improvements in system performance resulting from these additions.

4.1 The Task

For the experiment described in [6], 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. The story pointed out that the level of reliability between software and human performance was comparable and around 70%. 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 relying on their own predictions as well as the advice of the decision aid. The interface the participants were presented with is presented in the first 3 and the 6th rows of Figure 4. The task consitutes

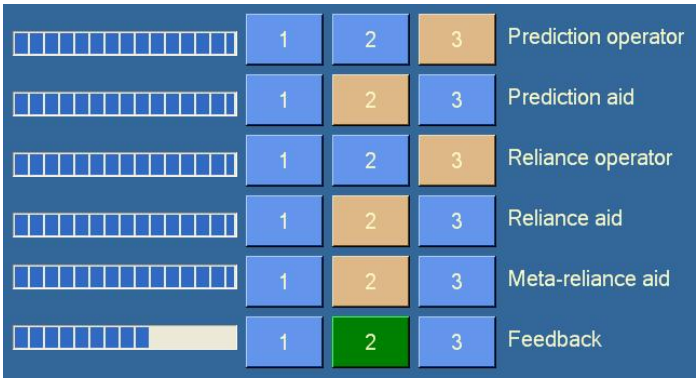

Fig. 4. An example interaction between the operator and the automated decision aid. The rows represent the different phases of the operator-aid task. For the current research, phases 4 and 5 were added to the task environment described in [6].

making a choice between 3 alternatives, as shown in each of the rows in the interface. In phase 1 the operator chooses, based on her own personal estimation of the pattern to be recognized. Then in phase 2 the machine chooses, with a pre-fixed average accuracy of 70%. Finally, in phase 3, the operator makes a reliance decision, by selecting the answer given in the first 2 phases by the agent she chooses to rely on. (The operator is free to choose a different answer altogether, but this happened only rarely in the experiments.) The last action of each trial consists of the feedback given by the system about which action was the correct one (phase 6), the corresponding button colored green if the operator's reliance decision was correct, and red if it was incorrect.

In order to support the operator in making reliance decisions the above operator-aid task was extended by adding 2 phases representing the aid's reliance (Sec. 3.2) and meta-reliance (Sec. 3.3) decisions. The next section details the aid's design in this respect.

4.2 Design of the Aid

In the original experiments, the aid did nothing more than provide an advice to the human operator. The enhancements to the aid's design were intended to provide the properties Reliance and Meta-reliance discussed in Section 3, to allow improvement upon the operator's Reliance Decision Making (RDM) in the form of Reliance Decision Making of the Machine (RDMM) and Meta-Reliance Decision Making of the Machine (Meta-RDMM).

Both RDMM and Meta-RDMM are based on a generic trust model [4] that allows the aid to estimate the operator's and the aid's abilities to make advice (task-related, prediction) and reliance decisions. The RDMM module makes the decision in phase 4 in Figure 4 ('Reliance Aid'), based on a comparison of the aid's trust in the operator's and the aid's own prediction abilities (phases 1 and 2). Like the operator in phase 3, the aid proposes in phase 4 the answer given in phases 1 and 2 by the agent it trusts most hightly, where trust refers to *prediction* capability. In case of disagreeing reliance decisions in phases 3 and 4, the aid chooses among the operator and the aid in phase 5, this time based on a comparison of its trust in the two agents' *reliance decision making* capabilities.

As mentioned above, the same basic trust model is used for both estimates (prediction and reliance decision making capabilities). Essentially, the respective abilities are modeled as random variables $0 \leq \theta_a^x \leq 1$, which are interpreted as the probabilities of each of the agents $a \in \{\text{operator, aid}\}$ making the correct decision $x \in \{\text{prediction, reliance}\}$. The aid uses Beta probability density functions (pdfs) over each of these 4 random variables to model its belief in each of the values of $\theta \in [0, 1]$ being the correct one. Based on the feedback obtained in phase 6, each of the answers given in phases 1 through 4 can be classified as 'success' or 'failure' depending on whether the operator and the aid, respectively, were correct or incorrect in their prediction and reliance decisions, respectively. At the end of each trial, the aid uses Bayes' rule to update each of its estimates given the newly obtained information from phase 6. The advantage of using a Beta pdf as a prior in Bayesian inference about a binomial likelihood (such as that of θ), is that the resulting posterior distribution is again a Beta pdf [12, 13].

In the next trial, the aid uses the new estimates about the agents' prediction abilities for RDMM in phase 4, and the estimates about the agents' reliance decision making abilities for Meta-RDMM in phase 5.

4.3 Experimental Results

The original experimental design and results are discussed in [6]. Here, we show to what extent the elaborations of the aid's design were able to enhance the system's overall performance. Table 1 shows these results. Each participant played two experiments of 101

	Operator-RDM	RDMM	Meta-RDMM
exp. 1	0.65	0.70	0.70
exp. 2	0.67	0.70	0.69
both	0.66	0.70	0.69

Table 1. Performance (percentage correct) of operator reliance decision making (Operator-RDM), RDMM, and Meta-RDMM. Per row, the differences between Operator-RDM and RDMM, and Operator-RDM and Meta-RDMM, are significant.

trials each. For each row, the improvements from operator reliance decision making (Operator-RDM) to RDMM, and from Operator-RDM to Meta-RDMM are significant. No significant difference in performance is found between RDMM and Meta-RDMM. There are no significant differences between experiment 1, 2, and both, for RDMM and Meta-RDMM. However, the differences between experiment 1, 2, and both, for Operator-RDM are significant. This means that, in our experiments, there was no measurable effect on performance of (Meta-)RDMM due to operator learning effects.

Our results indicate that the quality of the decision to rely on the prediction of either the operator or the aid is higher when it is made by RDMM than when it is made by human participants. When a computer would make reliance decisions based on RDMM it would outperform most human participants. However, it also became clear that in some situations humans make better reliance decisions than aids, and in others aids do. This means that reliance decision making completely done by the aid does not necessarily result in optimal performance. Meta-RDMM tries to take advantage of this and is based on the idea that the aid itself decides when to rely on RDMM and when to rely on the operator for reliance decision making (meta-reliance). Our results show that Meta-RDMM also outperforms human participants in reliance decision making, but (surprisingly) significant differences between RDMM and Meta-RDMM were not found.

5 Conclusion

The goal of augmented cognition is to extend the performance of human-machine systems via development and use of computational technology. In the context of the current work, performance can be improved when, like in human-human teams, both human and machine are able to assess and reach agreement on who should be trusted more and who should be relied on in what situation.

In this paper we showed that human reliance decisions are not perfect and reliance decision making can be augmented by computational technology. Our machine reliance decision making model outperforms human reliance decision making.

Now that we have our proof-of-concept in an abstract task, we intend to investigate how human-machine cooperation can be augmented in more complex and more realistic situations. We intend to focus on how models of trust and reliance can be practically used to adjust the level of autonomy of adaptive systems. We want to investigate in what domains this kind of support has an impact on the effectiveness of task performance, and how the magnitude of the impact depends on the task's and the domain's

characteristics. How serious are the conditions mentioned in section 3, both in terms of limiting the scope of application domains, and in terms of determining the effectiveness of our solutions. An important question is whether the properties of our abstract task environment are paralleled in real-world settings.

6 Acknowledgments

This research was partly funded by the Royal Netherlands Navy under program number V206 and by the Dutch government (SENTER) under project number TSIT2021.

References

1. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors **46** (2004) 50–80
2. Lee, J.D., Moray, N.: Trust, control strategies and allocation of function in human machine systems. Ergonomics **22** (1992) 671–691
3. Wiegmann, D.A., Rich, A., Zhang, H.: Automated diagnostic aids: the effects of aid reliability on user's trust and reliance. Theoretical Issues in Ergonomics Science **2** (2001) 352–367
4. Klos, T.B., La Poutré, H.: A versatile approach to combining trust values for making binary decisions. In: Trust Management. Volume 3986 of Lecture Notes in Computer Science. Springer (2006) 206–220
5. Maanen, P.-P. van, Dongen, K. van: 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), Utrecht, The Netherlands (July 2005) 168–77
6. Dongen, K. van, Maanen, P.-P. van: 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)
7. Parasuraman, R., Mouloua, M., Hilburn, B.: Adaptive aiding and adaptive task allocation enhance human-machine interaction. Automation Technology and Human Performance: Current Research and Trends **22** (1999) 119–123
8. Parasuraman, R., Riley, V.A.: Humans and automation: Use, misuse, disuse, abuse. Human Factors **39** (1997) 230–253
9. Moray, N., Inagaki, T., Itoh, M.: Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. Journal of Experimental Psychology: Applied **6** (2000) 44–58
10. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. International Journal of Human-Computer Studies **58** (2003) 697–718
11. Dongen, K. van, Maanen, P.-P. van: Designing for dynamic task allocation. In: Proceedings of the Seventh International Naturalistic Decision Making Conference (NDM7), Amsterdam, The Netherlands (2005)
12. D'Agostini, G.: Bayesian inference in processing experimental data: Principles and basic applications. Reports on Progress in Physics **66** (2003) 1383–1419
13. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Second edn. Chapman & Hall/CRC (2004)