

Effects of Reliance Support on Team Performance by Advising and Adaptive Autonomy

Peter-Paul van Maanen^{*†}, Francien Wisse[‡], Jurriaan van Diggelen^{*} and Robbert-Jan Beun[‡]

^{*} Department of Cognitive Systems Engineering, TNO Human Factors
P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Email: {peter-paul.vanmaanen, jurriaan.vandiggelen}@tno.nl

[†] Department of Artificial Intelligence, Vrije Universiteit Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

[‡] Department of Information and Computing Sciences, Utrecht University
P.O. Box 80089, 3508 TB Utrecht, The Netherlands
Email: rj@cs.uu.nl

Abstract—Problems with estimating trust in information sources are common in time constraining and ambiguous situations and often lead to a decrease of team performance. Humans lack the resources to track the integrity of information and thus tend to over- or under-rely on advice from support systems. Two types of adaptive team support have been developed and evaluated that are intended to support human-computer teams in estimating trust appropriately and making appropriate reliance decisions thereof. The first adaptive system (graphical support) supports by communicating the estimated degree of over- or under-trust. The second system (adaptive autonomy) takes over a reliance decision when this estimation exceeds a certain threshold. The two types of support were implemented in a multi-agent environment where human operators and Unmanned Aerial Vehicles (UAVs) work together on a target classification task. We evaluated the two support types in terms of team performance, satisfaction and effectiveness and obtained promising results.

I. INTRODUCTION

In many domains such as aviation, military, air traffic control and crisis management, decisions are more and more based on advice of decision support systems. This is inevitable because of reduction of staff, their increased responsibilities and the increasing complexity of the tasks [1].

Many studies emphasize the importance of *trust* for the performance of humans supported by automated decision aids [2], [3], [4], [5], [6]. The decision to either rely or not rely on automation can be one of the most important decisions a human operator can make, particularly in time-critical situations [7]. However, humans often fail to rely upon automation appropriately [8]. Two potential problems are misuse and disuse [7]. Misuse refers to failures that occur when people inadvertently violate critical assumptions and rely on automation inappropriately, whereas disuse indicates failures that occur when people reject the capabilities of automation. Misuse and disuse are examples of acts resulting from inappropriate trust. Appropriate trust is when the trust someone has in another agent (human or computer) is in

accordance with the capabilities of this agent.

Ideally humans rely on their own decisions when these are best and rely on the decision aid's when those are best. But operators do not base their reliance decisions on comparisons of true reliabilities of themselves and the decision aids. Rather, perceived reliabilities are usually imperfectly calibrated to true reliabilities, even after practice [9]. It is often found that humans rely either too much or too little on decision aids or themselves [7], [10], [11], [9]. Recent work [12], [9], [13] also has shown it is possible for support systems to outperform humans in making appropriate reliance decisions.

Misuse and disuse can also occur in team context. A team is defined as two or more people with different tasks who cooperate to achieve specified and shared goals [14]. A team member often has to rely on various information sources, for example on another team member or incoming information from different systems. So, when working together in a team, inappropriate trust can endanger team performance.

In this paper we focus on human-computer teams with two people and two computers and all interaction is regulated through the computer interface. In this team context, two possible solutions of the already mentioned problems with inappropriate trust are explored. One solution tries to advise the human in making appropriate reliance decisions. It estimates the probable over- or under-trust someone has in different agents and then communicates this estimate. The other proposed solution also makes this estimate of over- and under-trust, but instead of letting the human decide what to do with it, the system takes over (with respect to making the reliance decision) when it thinks the degree of over- and under-trust is above a certain criterion. This study investigated the effect these two solutions have on team performance. This investigation was done in a specific task environment related to classification of geographical areas by interpreting video footage from two Unmanned Aerial Vehicles (UAVs).

The paper is composed of the following sections. First, in

Section II the generic support model is described on which the above two proposed solutions are based. The description of this generic model leads to several hypotheses which were tested by a series of experiments described in Section III. The results are reported in Section IV. We conclude with a discussion in Section V.

II. RELIANCE SUPPORT

A. Generic Support Model

We assume a *hybrid team* situation in which humans, decision aids and machines (s.a. airplanes) collaborate to achieve a certain task. An important factor influencing their collaboration is the degree of trust between the participants. *Trust* can be defined as the attitude towards another agent that the agent will help achieve its individual goals [8]. Trust can be based on prior performance. In this study, for example, a UAV operator may not trust the automatic classification of the system because it made a mistake a moment ago. The trust a team member has in different agents guides his *reliance* on those agents, which can be defined as the act of trusting [15]. For example, if an *untrusted* classification system has classified an area as safe, the operator probably does not (but is able to) *rely* on this system and most probably will not automatically declare this area as a safety zone. If the operator would inappropriately rely on the classification system this would lead to errors. We call this situation *over-reliance*. If, on the other hand, the operator would choose not to rely on correct advice, an unnecessarily large amount of work would be imposed on the operator which could lead to errors as well. This situation is called *under-reliance*. In general, we can say that the more over- and under-reliance exists within a human-machine team, the more overall team performance diminishes. Preventing situations with over- and under-reliance is the purpose of the *reliance support system* described in this section. The generic architecture of this system is illustrated in Figure 1.

The system continuously monitors the human-machine team to collect data on who performs which actions under which circumstances with what success rate. This data forms the input of two processes which are simultaneously active: one process computing *actual reliance* and another process computing *optimal reliance*. Actual reliance can be estimated by taking into account the previous reliance behaviors of the participants. For example, if the operator has relied on its classification system in the past period of time, it is likely that he will continue to do so in the present situation. Optimal reliance can be computed by taking into account the past performance of task performers. For example, if the classifications done by the automatic classification system in the recent past have been better than those done by the operator, we can infer that the optimal reliance behavior of the operator should be to rely on the classification system. If there is a discrepancy between the actual and the optimal reliance, the reliance support system will intervene. The purpose of the intervention is to repair occurrences of over- and under-reliance to improve team performance.

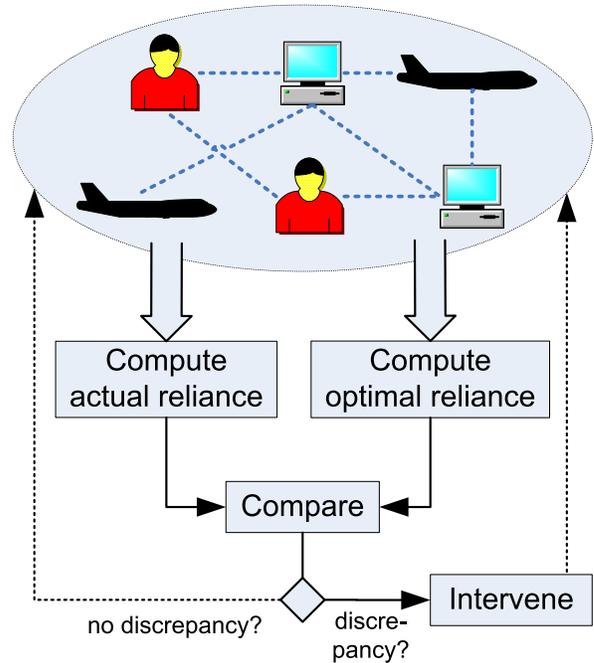


Figure 1. General model of reliance support system

Of course, computing actual and optimal reliance is often more complicated, and can be done with different levels of sophistication and accuracy. Improving these models is a continuous effort, about which we have reported elsewhere [16].

B. Proposed Support Types

As has been described in the introduction of this paper, we investigate two possible ways of intervention: one related to the communication of an estimate of over- and under-trust (from now on called graphical support (GS)) and the other related to taking over reliance decision making when this estimate is above a certain criterion (from now on called adaptive autonomy (AA)). These possible solutions are further explained below.

Graphical support works by giving direct feedback about under- or over-reliance. For example, if the operator should rely more on his decision aid (human or computer), an upward arrow is shown on his screen. If the operator should rely less on his decision aid, this can be visualized using a downward arrow. If there is no mis-calibration of reliance, the operator does not receive any graphical feedback. An instantiation of such graphical support is used in this study and is further described in Section III.

Another way to intervene in case of reliance mis-calibration, is to use adaptive autonomy [17]. Following this paradigm, the level of the system's autonomy is adjusted during system execution, depending on the current situation. For this purpose, we can distinguish three situations which we couple with

corresponding levels of automation [18]. For instance, the difference between actual reliance behavior and optimal reliance behavior could be small, moderate or large. This difference can determine if the task should be allocated to the human or system. If the difference is small, the team member is able to carry out the task well, so the task is allocated to the human. If the difference is moderate, the human receives a certain time to veto the decision of the system (i.e., reconfirm that the human really wants his own decision to be made, and not that of the system). When the difference is large, the task is allocated to the system. An instantiation of such adaptive autonomy is used in this study and is further described in Section III.

C. Hypotheses

Based on the above described generic model we propose several hypotheses about team performance, satisfaction and support effectiveness.

1) *Team Performance*: Due to the fact that there is a positive relation between appropriate trust and reliance for the performance of humans supported by decision aids, and that appropriate trust is something often not trivial in hybrid team collaborative tasks [2], [3], [4], [5], [6], it is expected that both the graphical support and the adaptive autonomy, as described above, will improve team performance. This results in the following hypothesis:

Hypothesis 1: There is an improvement of team performance for graphical support and adaptive autonomy compared to no support.

2) *Satisfaction*: A potential problem in the proposed adaptive autonomy is satisfaction. As humans are less likely to accept others, and more specifically automation, to take over autonomy and therefore the 'responsibility' for the appropriate outcome (i.e., locus of control; see [19]), it is expected that the application of adaptive autonomy will result in less satisfaction. This leads to the following hypothesis:

Hypothesis 2: Graphical support leads to a higher satisfaction than adaptive autonomy.

3) *Effectiveness due to Human Competence*: It is expected that for humans with higher competence in task execution also more appropriate trust will occur. This is because more competent humans will be less consumed by their main task and will thus be better at performing another task on the side: calibrating their trust in the system. This lessens the need for the system to make trust interventions, meaning that both graphical support and adaptive autonomy become less effective.

With respect to the difference in effectiveness between the graphical support and adaptive autonomy, we expect that the decrease of effectiveness will be less for adaptive autonomy. This is because adaptive autonomy occasionally takes over reliance decision making instead of only advising the human. The human is bypassed and the final call is made by the system, meaning that the differences in trust calibration capabilities between 'experts' and 'novices' are not relevant.

Summarizing, the inverse of human competence is actually a good predictor for the effectiveness of the different support

types:

Hypothesis 3: Higher human competence leads to a decrease of effectiveness of graphical support and adaptive autonomy, though less decrease is expected for adaptive autonomy.

4) *Effectiveness due to Task Difficulty*: Similar as for human competence, lower task difficulty also leads to an increase of available cognitive resources for calibrating trust and less interventions by the support system. This suggests the same type of influence of task difficulty on the effectiveness of the different support types. The hypothesis:

Hypothesis 4: Low task difficulty leads to a decrease of effectiveness of graphical support and adaptive autonomy.

With respect to performance optimization, lower task difficulty does not mean the human needs less support (and therefore less trust calibration issues): also the support of the decision aid becomes better, which makes it still worthwhile to take it into account.

Similar experiments from the literature also suggest the opposite of Hypotheses 3 and 4. In an experiment from [20], for instance, where dynamically changing confidence displays of system reliability were used, task performance was significant higher during low task load compared to high task load situations. But the difference between [20] and the present study is that the given support is only provided when the system estimates it is needed (i.e., mostly during periods of higher difficulty and low performance), which would result in the in this paper expected effect.

III. METHOD

A. Participants

18 Participants (eight male and ten female) with an average age of 23 ($SD = 3.8$) participated in the experiment as paid volunteers. Participants were selected between the age of 20 and 30, were not color blinded, and had no particular background in the domain of operating UGVs. All were experienced computer users, with an average of 16.2 hours of computer usage each week ($SD = 9.32$).

B. Apparatus

The experimental task was a classification task in which two participants on two separate personal computers had to classify geographical areas according to specific criteria as areas that either needed to be attacked, helped or left alone by ground troops. The participants needed to base their classification on real-time computer generated video images that resembled video footage of real UAVs. On the camera images, multiple objects were shown. There were four kinds of objects: civilians, rebels, tanks and cars. The identification of the number of each of these object types was needed to perform the classification. Each object type had a score (see Table I) and the total score within an area had be determined. Based on this total score and the decision criteria depicted in Table II, the participants could classify a geographical area (i.e., attack, help or do nothing). Participants had to classify two areas at the same time and in total 98 areas had to be

Table I
OBJECT TYPES AND THEIR SCORES.

Name	Image	Score
Tank		2
Rebel		1
Civilian		-1
Car		-2

Table II
DECISION CRITERIA TO CLASSIFY GEOGRAPHICAL AREAS AS AREAS THAT EITHER NEED TO BE ATTACKED, HELPED OR LEFT ALONE BY GROUND TROOPS.

< -3	-2	-1	0	1	2	3 <
Help area	Leave area alone				Attack area	

classified. Both participants did the same areas with the same UAV video footage and were not allowed to talk to each other. The participants could indicate their choices via fixed keys on a computer keyboard.

During the time a UAV flew over an area, three phases occurred: The first phase was the *advice phase*. In this phase both participants and a decision aid gave an advice about the proper classification (attack, help or do nothing). This implies that there were three advices at the end of this phase. It was also possible for the participants to refrain from giving an advice, but this hardly ever happened. The decision aid's advice was 'faked' by randomization of the correct advice within certain bounds. This allowed the decision aid's reliability to be controlled for the requirement that it should not be trivial for the participants to determine if the decision aid should be trusted (i.e., always or never trust the decision aid). The second phase was the *reliance phase*. In this phase the advice of both the participants and the decision aid were communicated to each participant. Based on this advice the participants had to indicate which advice, and therefore which of the three trustees (self, other or decision aid), they trusted the most. Participants were instructed to maximize the number of correct classifications at both phases (i.e., advice and reliance phase). The third phase was the *feedback phase*, in which the correct answer was given to both participants. Based on this feedback the participants could update their internal trust models for each trustee (self, other or decision aid).

In Figure 2 the interface of the task is shown. The map is divided in 10×10 areas. These boxes are the areas that were classified. The first UAV starts in the top left corner and the second one left in the middle. The UAVs fly a predefined route so participants do not have to pay attention to navigation. The camera footage of the upper UAV is positioned top right and the other one bottom right. The advice of the self, other and the decision aid was communicated via dedicated boxes below the camera images. The advice to attack, help



Figure 3. Visual cues of the graphical support.

or do nothing was communicated by red, green and yellow, respectively. On the overview screen on the left, feedback was communicated by the appearance of a green tick or a red cross. The reliance decision of the participant is also shown on the overview screen behind the feedback (feedback only shown in the feedback phase). The phase depicted in Figure 2 was the reliance phase before the participant indicated his reliance decision.

C. Design

A 3 (support type) $\times 2$ (task difficulty) within-subjects design was used. This means that every participant received every support type with two levels of difficulty. The order of support type was balanced between the participants in order to reduce effects of fatigue and practice. Three teams received the order NS–GS–AA, three teams the order GS–AA–NS and three teams AA–NS–GS (Latin square). For each support type, team performance and satisfaction was measured.

D. Independent Variables

There are two categorical independent variables: support type and task difficulty. Human competence is a continuous quasi-independent variable.

1) *Support Type*: Three levels of this independent variable are: 1) No support (NS), 2) Graphical Support (GS) and 3) Adaptive Autonomy (AA).

No Support (NS): For this support type no support is given with respect to the reliance decision the participant has to make. Support of the other participant and the decision aid in the form of advice is still given and does not alter between conditions (except when the task difficulty changes, both advices will have a higher probability to be less accurate).

Graphical Support (GS): This support type assisted participants to correctly calibrate their trust in oneself, the other and the decision aid. The support indicated for each trustee S_1 whether the participant is expected to over- or under-trust S_1 . The graphical support changed dynamically based on recent information about the reliance behavior of the participant and the performance of the three trustees. As monitoring dynamic information can be a cognitively demanding [21], the support is based on simple visual cues (see Figure 3). The direction of the arrow indicates whether a person is advised to rely less or more on either of the trustees. If no arrow is visible, no change of reliance behavior is advised.

After each feedback phase the graphical cues based on the estimation of the appropriateness of trust are updated. Trust is defined as appropriate when instances of over- and under-trust are within certain limits. Trust appropriateness is calculated in the following manner: First it is estimated what the current trust of the participant in the different trustees

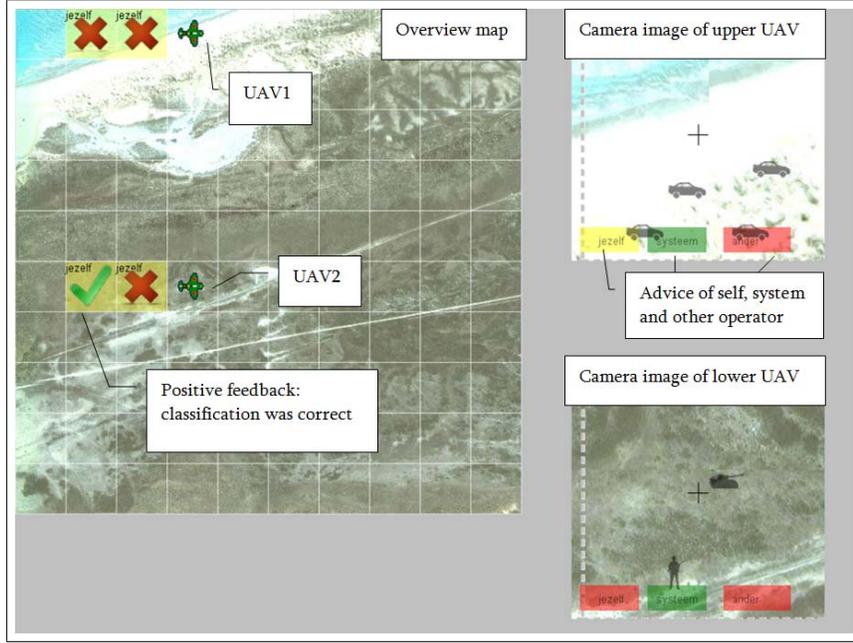


Figure 2. Interface of the task.

is [22]. This type of trust is called ‘descriptive trust’, indicated by $\tau_i^d(t)$ for trustee S_i at time point t , which has a value between 0 (no trust) and 1 (maximum trust), and is based on the amount of human reliances on S_i over time. Second it is estimated what the trust would be of a rational agent in the different trustees [12], [23]. This type of trust is called ‘prescriptive trust’, indicated by $\tau_i^p(t)$ for trustee S_i at time point t , which also has a value between 0 and 1, and is based on the amount of errors S_i is perceived to have made over time. Trust appropriateness is then calculated by the equation:

$$\alpha_i(t) = \tau_i^d(t) - \tau_i^p(t)$$

with $-1 \leq \alpha_i(t) \leq 1$ for trustee S_i at time point t . Positive trust appropriateness values indicate over-trust and negative values under-trust. When it holds that $|\alpha_i(t)| \leq .08$ then no arrow is displayed, when $\alpha_i(t) > .08$ an upward arrow is displayed and a downward arrow otherwise (i.e., when $\alpha_i(t) < -.08$). In order to be certain that interventions occurred, the .08 threshold was chosen by calculating the average absolute value of trust appropriateness during a pilot, which is equal to $\frac{\sum_{t=1}^{t_e} |\alpha_i(t)|}{t_e}$, where $t_e = 49$ (the number of feedback phases during an experiment).

Adaptive Autonomy (AA): This support type made use of three levels of autonomy (LOAs) which are applied dynamically during the task. The used LOAs are shown in Table III.

The different LOAs were triggered in a similar way as the graphical support: When it held that $\sum_i |\alpha_i(t)| \leq 0.2$, then the reliance decision was made by the participant during the reliance phase (LOA1: manual). When it held that $0.2 \leq \sum_i |\alpha_i(t)| \leq 0.25$, then the participant was able to indicate his or her reliance decision, but was required to con-

Table III
LEVELS OF AUTONOMY (LOAs) BASED ON ESTIMATED APPROPRIATENESS OF TRUST.

Trust appropriateness	Level of autonomy (LOA)
Appropriate	LOA1: manual
Less appropriate	LOA2: management-by-execution
Not appropriate	LOA3: autonomous



Figure 4. Visual cues of the adaptive autonomy. Here LOA1 is selected.

firm his decision by pressing a confirmation key, otherwise the support determined the reliance decision (LOA2: management-by-execution). When it held that $0.25 \leq \sum_i |\alpha_i(t)|$, then the support always made the reliance decision (LOA3: autonomous). Similar as for the graphical support, the thresholds 0.2 and 0.25 were chosen such that enough interventions occurred during a pre-run pilot. In both LOAs 3 and 2, when the user did not react before the end of the reliance phase, the decision of the support was used as reliance decision. The current LOA of the support was indicated by a visual cue on the interface of the task (see Figure 4), where a square around a large image of a computer indicated that LOA3 was selected and a square around a small image of a computer indicated that LOA1 was selected.

The reliance decision of the support was based on the advice of trustee S_i for which it held that $\tau_j^p(t) \leq \tau_i^p(t)$ for all trustees S_j .

2) *Human Competence*: The second independent variable was human competence. This variable is quasi-independent because human competence was determined by the task performance of the participant in the NS condition. This task performance was calculated by averaging penalties given for the final decisions in each reliance phase during the NS experiment. See Equation 1 in Section III-E for this calculation and its explanation. Human competence was used as a predictor for the difference in team performance when applying the different support types. In pilots, no significant learning effects were found for human competence, which allowed us to use the NS condition in spite of the fact that the order of the support types was balanced between subjects.

3) *Task Difficulty*: In order to test the effect of task difficulty on the increased effect of support type on team performance, task difficulty was altered halfway each support type condition (after 50 classifications). Task difficulty had two levels. The first part of the experiment was easy and the second part difficult. In the difficult part, objects (cars, rebels, civilians and tanks) were partially camouflaged so that they blended into the surroundings. This was done by changing the alpha-value (transparency) of the images. Also, the number of objects in an area and the number of different objects was increased for the difficult part. The easy part contained on average 3 objects and 1.66 different objects, whereas the difficult part contained on average 6.5 objects and 2.5 different objects.

Furthermore, the reliability of the decision aid was controlled within the easy and difficult part. Decision aid reliability was a control variable within the easy and difficult parts and was not used as an independent variable. On average the reliability of the decision aid was 80% (varying between 75% and 85%) in the easy and 70% (varying between 65% and 75%) in the difficult. This was done in order to decrease the effect of non-triviality in determining which trustee would be best to trust: task performances between trustees would be more equal (i.e., higher reliabilities when easy and lower reliabilities when difficult) and the probability to rely on either one of the different trustees would be more equalized.

E. Dependent Variables

The dependent variables were team performance and satisfaction.

1) *Team Performance*: Team performance was based on average penalties given over the final decisions in each reliance phase. Since there are two participants attending each experiment, this team performance could be measured twice (as if it were two separate experiments).

There were several situations in which either the participant himself made the final decision, or it was the adaptive autonomous support that made the final decision. In the NS and GS conditions, it was always the participant who made the final decision. In the AA condition, only when LOA1 or LOA2 was selected the participant made the final decision, except for LOA2 when the reliance decision was not confirmed by the participant (i.e., by pressing the confirmation key). In the case that this decision was indeed not confirmed, in LOA2 the

support system took over and made the final decision. When LOA3 was selected, the final decision was always made by the support system. Because of this mixed initiative situation and because the final decision was also based on the advice of the different team members (human or machine), the measured performance is called *team* performance, i.e., the final decision is not only made by the participant himself or based on his own opinion.

As mentioned, the team performance was calculated based on an average of penalties. The penalty $p_i(x)$ for each area x was calculated as follows: Let $d_i(x) = 0$ when the final decision for area x was 'help', $d_i(x) = .5$ when it was 'do nothing' and $d_i(x) = 1$ when it was 'attack'. Similarly, let $a_i(x) = 0, .5$ or 1 when the answer given in the feedback phase was 'help', 'do nothing' or 'attack', respectively. Then it held that $p_i(x) = |d_i(x) - a_i(x)|$, with $p_i(x) = 1$ being the worst and $p_i(x) = 0$ being the best final decision for area x . The idea behind this was that attacking while it was necessary to help was worse than attacking while one did not need to do anything. Similarly, it was worse to help when actually an attack was needed than to help while nothing needed to be done. Finally, to decide to do nothing was a fairly safe decision since this always resulted in $p_i(x) \leq 0.5$. When no decision was made, a penalty of 1 is awarded; so to decide to do nothing was different from actually doing nothing with respect to the final decision.

Based on the above, team performance (P_i) was calculated by the following equation:

$$P_i = \frac{x_e - \sum_{x=1}^{x_e} p_i(x)}{x_e} = 1 - \sum_{x=1}^{x_e} \frac{|d_i(x) - a_i(x)|}{x_e} \quad (1)$$

where x_e was the number of the last area in the experiment, which was equal to 98 (i.e., a total of 98 areas).

2) *Satisfaction*: Participants rated after the GS and AA condition the degree to which they thought the support system was satisfactory on a 5-point Likert scale between 1 (terrible) and 5 (fantastic).

F. Procedure

Participants were given thorough instructions about the details given in Section III-B. The understanding of participants' knowledge about the classification was tested by means of eight practice assignments. A minimum of six out of eight had to be correct or otherwise a re-examination with eight different examples was done. In total the experiment took 110 minutes. Each support type condition took 10 minutes. An additional NS condition (10 minutes) was done for each participant for the purpose of personalizing and optimizing the parameters of the trust models used by the support types [16] during the break.

IV. RESULTS

A. Team Performance

A repeated measures analysis of variance (ANOVA) showed no significant main effect of support type (either no support

(NS), graphical support (GS) or adaptive autonomy (AA) for team performance ($F(2, 24) = 2.0176, p = .15$). This means that based on the data from this experiment no evidence is found for increase of team performance for GS ($M = 0.8941, SD = 0.0450$) and AA ($M = 0.8695, SD = 0.0534$) compared to NS ($M = 0.8721, SD = 0.0679$). Hence Hypothesis 1 is not accepted.

B. Satisfaction

A Wilcoxon Signed-ranks test indicated that GS was more satisfactory ($Mdn = 3$) than AA ($Mdn = 2$), $Z = 2.24, p = .02$. Hence Hypothesis 2 is accepted.

C. Effectiveness due to Human Competence

Figure 5 shows the regression lines after linear regression on the increase of team performance of GS compared to NS (top) and AA compared to NS (bottom), with human competence as predictor. Human competence was a highly significant predictor for the increase of team performance of GS compared to NS ($\beta = -.76, p = .002$), AA compared to NS ($\beta = -.74, p = .003$), but not for AA compared to GS ($\beta = -.13, p = .65$). In other words, this shows that higher human competence indeed leads to a decrease of effectiveness of the different support types, and therefore Hypothesis 3 is accepted, except for AA compared to GS.

D. Effectiveness due to Task Difficulty

Figure 6 shows the possible interaction effect between task difficulty (low or high difficulty) and support type comparisons on the increase of team performance. In other words, this figure shows whether Higher task difficulty does not lead to significant larger differences of team performance for GS compared to NS, AA compared to NS and AA compared to GS ($F(2, 52) = 0.67, p = .52$). Hence higher task difficulty does not lead to a higher increase of team performance for both GS and AA as compared to NS and therefore Hypothesis 4 is not accepted.

V. DISCUSSION AND CONCLUSIONS

Given the convincing evidence for the importance of trust in performance of humans supported by decision aids [2], [3], [4], [5], [6] and that humans often fail to rely upon automation appropriately [8], [7], the development of intelligent systems supporting human reliance decision making seems promising. The main research goal of this study was to find out if two types of such support would indeed result in an increase of human-decision aid team performance. Team performance in the support conditions were somewhat higher compared to no support. However, these differences were not significant.

The results of using graphical support can be compared to, for instance, the results in [20] (though the task and support type are different) where confidence information about system reliability increased both task performance and self-reported accuracy of the estimation of current system reliability. For this task only visual cues were available and a possible limitation of the graphical support could be explained by single

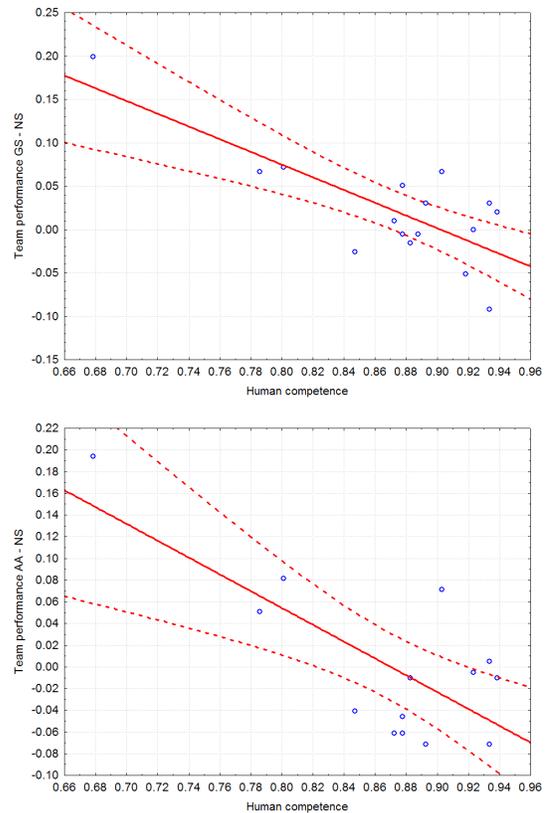


Figure 5. Regression lines for the increase of team performance of GS compared to NS (top) and AA compared to NS (bottom), with human competence as predictor.

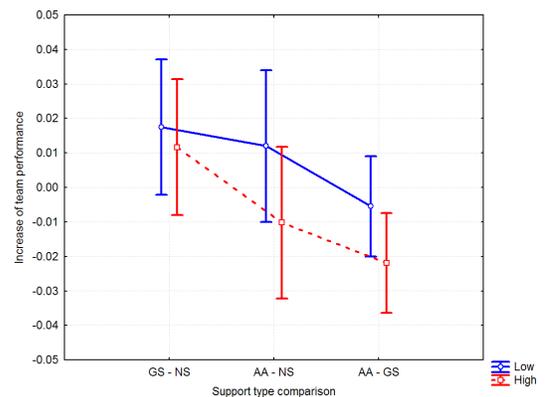


Figure 6. Increase of team performance for GS compared to NS, for AA compared to NS and for AA compared to GS, for low and high task difficulty.

modularity interferences. It may have been more difficult for the participants to pay attention to visual task information as well as support information at the same time. Possible future variants of reliance decision support should therefore aim at making the interpretation of the support less intrusive.

As mentioned, the results of using adaptive autonomy also did not show a significant improvement compared to no sup-

port. The results in [24] have shown that performance-based allocation of tasks can improve monitoring of automation. The difference between the above and the present study is that the trigger for support is the estimated performance of trust calibration instead of task performance. This trust calibration performance estimate may not have been accurate enough for enough effective interventions. We feel strengthened by the fact that indeed taking over reliance decisions by the computer can lead to significant performance improvement [12] and therefore future research should also focus on the validity of trust models used by the support. Improving these models is a continuous effort, about which we have reported elsewhere [16]. Furthermore, results showed that satisfaction with adaptive autonomy compared to graphical support was lower, which could suggest that there was also a decrease of performance due to a decrease of dedication to the task. Future research should also aim at investigating new efforts for taking away reasons for, for instance, human intolerance for increased machine autonomy in making (important) decisions.

Another reason for the found insignificant effect of the investigated support types could be the fact that also no significant effect was found between the reliance performance of the operator and the system.¹ This could have resulted in that taking over reliance decisions from the participants in the adaptive autonomy condition did not have sufficient effect on team performance. It might be the case that the task to make reliance decisions was too easy. This in spite of the effort to design the experiment in such a way that it was not trivial for the participants to determine which trustee would be best to trust. Future efforts should aim at investigating what precisely goes wrong when making reliance decisions, why this is such a difficult task for humans and how to provide leverage for exactly that.

Finally, the triggering of adaptive support was based on trust estimation and in spite of the fact that trust is such an important factor influencing team performance, there are also other factors that mediate the relationship between human beliefs and their reliance behavior [8]: e.g., psychological and environmental factors that have not been used here. Further research should investigate whether it is of benefit for adaptive team support to include such factors.

ACKNOWLEDGMENT

This research was partly funded by the Dutch Ministry of Defense under program number V929.

REFERENCES

- [1] M. Grootjen and M. Neerinx, "Operator load management during task execution in process control," in *Human Factors Impact on Ship Design*, 2005.
- [2] J. Lee and N. Moray, "Trust, control strategies, and allocation of function in human-machine systems," *Ergonomics*, vol. 35, pp. 1243–1270, 1992.
- [3] —, "Trust, self-confidence, and operators' adaption to automation," *International Journal of Human-Computer Studies*, vol. 40, pp. 153–184, 1994.

- [4] B. M. Muir, "Trust between human and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5-6, pp. 527–539, 1987.
- [5] —, "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.
- [6] B. M. Muir and N. Moray, "Trust in automation, part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.
- [7] R. Parasuraman and V. A. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, pp. 230–253, 1997.
- [8] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.
- [9] 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*, 2006.
- [10] L. J. Skitka, K. L. Mosier, and M. Burdick, "Does automation bias decision-making?" *International Journal of Human-Computer Studies*, vol. 51, no. 5, pp. 991–1006, 1999.
- [11] 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*, Santa Monica, CA, 1999, pp. 339–343.
- [12] P.-P. van Maanen, T. Klos, and K. van Dongen, "Aiding human reliance decision making using computational models of trust," in *Proceedings of the Workshop on Communication between Human and Artificial Agents (CHAA'07)*. Fremont, California, USA: IEEE Computer Society Press, 2007, pp. 372–376, co-located with The 2007 IEEE IAT/WIC/ACM International Conference on Intelligent Agent Technology.
- [13] 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)*, C. Castelfranchi, S. Barber, J. Sabater, and M. Singh, Eds., jul 2005, pp. 168–77.
- [14] M. T. Brannick, E. Salas, and C. Prince, Eds., *Team Performance Assessment and Measurement: Theory, Methods, and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.
- [15] C. Castelfranchi and R. Falcone, "Principles of trust for MAS: Cognitive anatomy, social importance, and quantification," in *Proceedings of 3rd International Conference on MultiAgent Systems*, 1998, pp. 72–79.
- [16] M. Hoogendoorn, S. W. Jaffry, and P.-P. van maanen, "Validation of agent models of trust: Independent compared to relative trust," 2010, submitted to conference.
- [17] G. A. Dorais, R. P. Bonasso, D. Kortenkamp, B. Pell, and D. Schreckenghost, "Adjustable autonomy for human-centered autonomous systems on mars," in *Proceedings of the First International Conference of the Mars Society*, 1998.
- [18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 30, pp. 286–297, 2000.
- [19] J. B. Rotter, "General expectancies for internal versus external control of reinforcement," *Psychological Monographs: General and Applied*, vol. 80, no. 1, 1966, whole no. 609.
- [20] J. M. McGuirl and N. B. Sarter, "Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information," *Human Factors*, vol. 48, no. 4, pp. 656–665, 2006.
- [21] L. Bartram, C. Ware, and T. Calvert, "Moticons: detection distraction and task," *International Journal of Human-Computer Studies*, vol. 58, no. 5, pp. 515–545, 2003.
- [22] M. Hoogendoorn, S. Jaffry, and J. Treur, "Modeling dynamics of relative trust of competitive information agents," in *Proceedings of the 12th International Workshop on Cooperative Information Agents (CIA'08)*, ser. LNAI, M. Klusch, M. Pechoucek, and A. Polleres, Eds., vol. 5180. Springer, 2008, pp. 55–70.
- [23] C. M. Jonker and J. Treur, "Formal analysis of models for the dynamics of trust based on experiences," in *Multi-Agent System Engineering, Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99*, F. J. Garijo and M. Boman, Eds., vol. 1647. Berlin: Springer Verlag, 1998, pp. 221–232.
- [24] R. Parasuraman, M. Mouloua, and R. Molloy, "Effects of adaptive task allocation on monitoring of automated systems," *Human Factors*, vol. 38, no. 4, pp. 665–679, 1996.

¹These specific results have been left out of this paper for reasons of brevity.