The Effects of the Adaptability and Reliability of Automation on Performance, Stress and Workload

Grant S. Taylor & James L. Szalma, Ph.D.
University of Central Florida

One way in which non-human agents manifest in complex systems is through the use of adaptive automation. In this case, the degree of involvement of a computer-based agent can be controlled dynamically as a function of environmental conditions and operator state. The current study examined the effects of specific aspects of adaptive automation, reliability and adaptability (the ability of the system to adapt appropriately to changes in task demand), on human performance, stress, and workload.

INTRODUCTION

Advancements in hardware and software technology have led to the development and implementation of autonomous and semi-autonomous non-human agents. In many instances the non-human agent is computer-based automation of system functions that permit the human team member to direct attention to critical mission tasks ill-suited for automation. Adaptive automation (AA) permits dynamic role assignment between the human and non-human agents as a function of changing conditions, including task demands and operator cognitive and affective state (Hancock & Chignell, 1988; Kaber & Endsley, 2004; Sheridan, 2002). Advances in adaptive automation promise to improve the human-agent synergy across a variety of environmental conditions, thereby improving overall system effectiveness. In adaptive automation, the human and the agent (automation) share task responsibilities at different points in a work cycle, depending on both environmental conditions and operator functional state (e.g., see Scallen, Hancock, & Duley, 1995).

Critical to successful human-agent interaction using AA is in determining how operator performance and stress state vary as a function of changes in level of automation (LOA; see Parasuraman, Sheridan, & Wickens, 2000) and task demands, and the task and environmental variables that influence how operators appraise and respond to the non-human agent. However, these technological advancements have introduced new problems in operation and coordination of human and agent activity. In AA, a particularly important issue is how operators respond to changes in automation and how such changes interact with task demand to influence overall human-agent system performance. In this study the effects of automation reliability and adaptability (i.e., whether the level of automation is increased when task demand increases and is decreased when task demand decreases) on performance, stress, and workload are examined in hopes of better understanding how aspects of AA impact the operator’s functional state.

METHOD

Participants

A total of 162 university students (age: M = 19.8, SD = 2.59) participated in the experiment, of which 52 were male and 110 female. Participants chose to receive either course credit or $16 for their participation in the 2-hour study. All participants received an additional cash bonus based on their performance, which was generally $4 to $6.

Procedure

Scenario. Participants were presented with the scenario that a terrorist group had infiltrated a government office building in the United States. An initial reconnaissance mission would be necessary to determine the location and number of terrorists, civilians, friendly forces, and weapons within the building. Given the danger involved in this mission, a team of unmanned ground vehicles (UGVs) would be sent into the building to transmit surveillance video to an operator.

Task. Participants were instructed that the movement of the UGVs was fully automated, so they had only to monitor their “videos feeds” (the videos were actually prerecorded and presented in a fixed random order). Participants viewed a total of 420 short video clips (10 seconds each, 2.5” high x 3.5” wide) of the simulated view of a UGV maneuvering through different rooms.

The participant’s task was to visually scan the video of each room as the UGV moved through them. After the UGV passed through a room, the participant was asked to indicate whether they saw a terrorist, friendly soldier, civilian, IED (improvised explosive device), or none of the above, no single video contained more than one of these items. Participants had 7 seconds to make a response before the next video began to play. During the response time, the automation was either at a high or low level, depending on the condition and block (described below). For the low automation level the participant was provided with a list of all possible signals of interest from which to select their response (level 2 on the level of automation scale proposed by Parasuraman, Sheridan, & Wickens, 2000). During high automation the aid provided a suggestion by highlighting one of these responses, but still left the final decision of which to select to the participant (level 4 automation).

The videos were broken up into a total of four 10-minute blocks of 35 trials each. Task demand was manipulated such that participants were required to monitor video displays from 2 UGVs simultaneously for each trial in blocks 1 and 3 (low task demand) and monitor displays from 4 UGVs for each trial in blocks 2 and 4 (high task demand).

Manipulations. Two independent variables were manipulated between subjects: reliability and adaptability.
Reliability was the accuracy of the suggestions made by the aid in the high automation trials. Each participant was randomly assigned to receive either high or low reliability automation. The high reliability aid was correct 95% of the time, while the low reliability aid was correct 75% of the time. These values were developed based on pilot data to be one standard deviation above and below the average human correct detection rate. Participants had no direct knowledge of the reliability of the system, aside from being instructed that “the automation may make mistakes as it is based on a computer vision algorithm.”

Each participant was also randomly assigned to receive either matched or mis-matched automation. The matched automation would provide low level automation when task demands were low in blocks 1 and 3, and high level automation when task demand was high in block 2 (block 4 also had a high task demand, but was treated as a separate condition as described below). The mis-matched automation provided the opposite pattern: a high level of automation when task demands were low and a low level of automation when task demands were high.

All participants were given the option of having the automated aid on or off for block 4 (they were able to control whether it was on or off throughout this block). All participants received high reliability automation for this block when they chose to have the aid engaged. Participants were also told that they would receive a cash bonus based on their performance in the fourth block of $0.05 for each correct response and a deduction of $0.01 for each incorrect response, which led to a maximum bonus of $7. Most participants received between $4 and $6.

Measures. Measures of perceived workload (NASA-TLX; Hart & Staveland, 1988) and mood state (DSSQ; Matthews et al., 1999) were administered to participants after each block of trials, with an additional administration of the DSSQ before the first block to be used as a baseline measurement. Participants answered all questionnaires directly on the same computers used for the experimental task.

RESULTS

Percent Correct

A repeated measures ANOVA was used to test the combined effects of block (1, 2, and 3, repeated measures), reliability (low and high, between subjects) and adaptability (matched and mis-matched automation, between subjects) on the performance of the task (measured by the percent of objects correctly identified in the videos). Due to the differences in the characteristics of the automation experienced in block 4, the scores from this block were analyzed separately.

Results (see Figure 2) showed a significant main effect for block, $F(2,398) = 32.688, p < .001$, such that performance in block 3 ($M = 83.9\%$) was significantly better than performance in both blocks 1 ($M = 80.3\%$) and 2 ($M = 80.8\%$), suggesting that participants became better at the task over time. There was also a significant main effect for reliability, $F(1,199) = 12.746, p < .001$, such that the high reliability group ($M = 83.2\%$) performed significantly better than the low reliability group ($M = 80.1\%$), showing that the difference in the reliability of the automation was enough to significantly impact performance.

Significant interactions were found between block * reliability, block * adaptability, and block * reliability * adaptability. The block * reliability interaction, $F(2,398) = 4.455, p = .012$, showed that the high reliability group performed better in block 3 ($M = 85.1\%$) than block 1 ($M = 81.3\%$), while the low reliability group performed better in block 3 ($M = 82.7\%$) than both blocks 1 ($M = 79.2\%$) and 2 ($M = 78.4\%$). The block * adaptability interaction showed that the mis-matched group performed better in blocks 1 ($M = 82.8\%$) and 3 ($M = 85.6\%$) than in block 2 ($M = 78.7\%$), and better in block 3 than block 1. The matched group performed better in blocks 2 ($M = 82.6\%$) and 3 ($M = 82.2\%$) than block 1 ($M = 77.8\%$). This is to be expected, as the mis-matched group received assistance with the relatively easier blocks 1 and 3, and the matched group show the same practice effect as that shown in the main block effect, with performance on block 2 improved due to the automation.

The block * reliability * adaptability interaction, $F(2,398) = 40.196, p < .001$, showed that for both matched and mis-matched groups, when reliability was low, performance for block 3 was better than both blocks 1 and 2. When reliability was high, the matched group performed better in block 2 than both blocks 1 and 3, and better in block 3 than 1, the mis-matched group performed better in blocks 1 and 3 than in block 2. These results show that the low reliability automation has little impact on performance, regardless of whether it is matched properly to task demand or not, and the high reliability automation improves performance whenever it is engaged.

No significant effects were found for performance in block 4.

![Figure 1: Legend for all graphs](image)

![Figure 2: Percent Correct](image)
Stress

The Dundee Stress-State Questionnaire (DSSQ) reports a total of 10 components of stress: Energetic Arousal (alertness or sluggishness), Tense Arousal (nervousness or relaxation), Hedonic Tone (general feelings of happiness/cheerfulness), Intrinsic Task Motivation, Self-Focused Attention (self-reflection), Self-Esteem, Concentration, Confidence and Control, Task-Relevant Cognitive Interference (worry about task performance), and Task-Irrelevant Cognitive Interference (worry about personal concerns). A repeated measures ANOVA was conducted on each of these scales, with scores from the pre-task measure and those collected after blocks 1, 2, and 3 as the repeated measures factor, and reliability and adaptability as between-subject factors. Due to the task differences in block 4, the effects of reliability and adaptability on block 4 stress were tested in a separate ANOVA for each component.

Energetic Arousal. A main effect for block was found, F(3,471) = 31.535, p < .001, such that Energetic Arousal for block 2 (M = .274) was lower than the pre-task state (M = .164), block 1 (M = .326), and block 3 (M = .303). No significant effects were found in block 4.

Tense Arousal. A main effect for block was found, F(3,471) = 30.817, p < .001, such that Tense Arousal for blocks 1 (M = .401), 2 (M = .562), and 3 (M = .233) were all higher than the pre-task state (M = .027). Block 2 was also higher than blocks 1 and 3. A main effect for reliability was found for block 4, F(1,157) = 4.045, p = .046, such that the low reliability group experienced greater Tense Arousal (M = .493) than the high reliability group (M = .215) during block 4.

Hedonic Tone. A main effect for block was found, F(3,471) = 30.817, p < .001, such that Hedonic Tone for blocks 1 (M = -.295), 2 (M = -.443), and 3 (M = -.445) were all lower than the pre-task state (M = .208). No significant effects were found in block 4.

Motivation. A main effect for block was found, F(3,471) = 21.016, p < .001, such that Motivation for blocks 1 (M = .448), 2 (M = .384), and 3 (M = -.092) were all lower than the pre-task state (M = .786). Blocks 1 and 2 were also higher than block 3. No significant effects were found in block 4.

Self-Focused Attention. A main effect for block was found, F(3,471) = 157.788, p < .001, such that Self-Focused Attention for blocks 1 (M = -.868), 2 (M = -1.034), and 3 (M = -1.035) were all lower than the pre-task state (M = .150). Block 1 was also higher than blocks 2 and 3. No significant effects were found for in block 4.

Self-Esteem. A main effect for block was found, F(3,471) = 61.071, p < .001, such that Self-Esteem for blocks 1 (M = .696), 2 (M = .792), and 3 (M = .871) were all higher than the pre-task state (M = .263). Block 3 was also higher than blocks 1 and 2. No significant effects were found in block 4.

Concentration. A main effect for block was found, F(3,471) = 16.197, p < .001, such that Concentration for blocks 1 (M = .744) and 2 (M = .688) were higher than the pre-task state (M = .498). Blocks 1 and 2 were also higher than block 3 (M = .367). A main effect for reliability, F(1,157) = 8.897, p = .003, was found such that the Concentration was higher for the high reliability group (M = .749) than the low reliability group (M = .399). A reliability * match interaction was also found, F(1,157) = 6.864, p = .010, such that there was no effect for adaptability under the high reliability condition, but the matched automation group experienced greater Concentration (M = .619) than the mis-matched group (M = 179) for the low reliability condition (Figure 3). No significant effects were found in block 4.

Confidence and Control. A main effect for block was found, F(3,471) = 28.717, p < .001, such that Confidence and Control for blocks 1 (M = 1.218), 2 (M = .944) and 3 (M = 1.003) were all lower than the pre-task state (M = 1.690). Block 1 was also higher than block 2. No significant effects were found in block 4.

Task-Relevant Cognitive Interference. A main effect for block was found, F(3,471) = 17.882, p < .001, such that Task-Relevant Cognitive Interference for blocks 1 (M = .419), 2 (M = .491), and 3 (M = .356) were all lower than the pre-task state (M = .270). Block 2 was also lower than block 3. A block * match interaction was found, F(3,471) = 2.851, p = .037, such that Task-Relevant Cognitive Interference for both matched and mis-matched groups was higher for the pre-task state than blocks 1, 2, and 3, but Task-Relevant Cognitive Interference was higher for block 3 than for the matched group only. A reliability * match interaction was also found, F(1,157) = 12.665, p = .027, such that there was no effect for reliability within the matched condition, but the low reliability group experienced more Task-Relevant Cognitive Interference (M = .003) than the high reliability group (M = .461) within the mis-matched automation condition. A main effect for reliability was found for block 4, F(1,157) = 3.717, p = .016, such that the low reliability group (M = .293) experienced more Task-Relevant Cognitive Interference than the high reliability group (M = .597) (Figure 4).
Workload

The NASA-TLX workload measure yields scores for 6 separate components of workload: Mental Demand, Temporal Demand, Effort, Frustration, Perceived Performance, and Physical Demand, each rated on a scale from 0 to 500. These components are combined into a single Global Workload measure, which ranges from 1 to 100. The NASA-TLX was administered after each block of trials, and so a repeated measures ANOVA was conducted to determine the combined effects of automation adaptability, reliability, and experimental block on workload. As with performance and stress, measures related to block 4 were analyzed separately.

**Mental Demand.** A main effect for block was found, $F(2,342) = 64.921, p < .001$, such that Mental Demand for block 1 ($M = 256.9$) was higher than block 3 ($M = 211.8$), and block 2 ($M = 314.9$) was greater than both blocks 1 and 2.

**Temporal Demand.** A main effect for block was found, $F(2,342) = 41.961, p < .001$, such that Temporal Demand was greater for block 1 ($M = 157.5$) than block 3 ($M = 126.1$), and block 2 ($M = 210.7$) was greater than both blocks 1 and 3. A reliability * match interaction, $F(1,171) = 5.770, p = .017$, was found such that when reliability was high, the mismatched group ($M = 179.8$) was greater than the matched group ($M = 135.0$), while there was no effect of adaptability when the reliability was low (Figure 5).

**Effort.** A main effect for block was found, $F(2,342) = 35.339, p < .001$, such that Effort was greater for block 2 ($M = 209.1$) than blocks 1 ($M = 159.0$) and 3 ($M = 140.0$).

**Frustration.** A main effect for block was found, $F(2,342) = 4.487, p = .012$, such that Frustration was greater for block 2 ($M = 102.5$) than block 1 ($M = 77.0$).

**Perceived Performance.** A main effect for block was found, $F(2,342) = 7.065, p = .001$, such that Perceived Performance was greater for block 3 ($M = 140.0$) than block 2 ($M = 107.6$).

**Physical Demand.** A main effect for block was found, $F(2,342) = 6.640, p = .001$, such that Physical Demand was greater for blocks 2 ($M = 15.60$) and 3 ($M = 15.98$) than block 1 ($M = 4.39$). An effect for reliability was found in block 4, $F(1,171) = 4.899, p = .028$, such that the low reliability group experienced greater Physical Demand ($M = 28.19$) than the high reliability group ($M = 11.06$).

**Global Workload.** A main effect for block was found, $F(2,342) = 81.344, p < .001$, such that Global Workload was greater for block 1 ($M = 52.06$) than block 3 ($M = 48.11$), and block 2 ($M = 64.03$) was greater than both blocks 1 and 3 (Figure 6).

**DISCUSSION**

Results of the current study indicate that the performance, workload, and stress associated with performing tasks with or without an adaptive aid change as a function of time, the level of demand, and the degree to which the adaptive automation is well calibrated to task demand (i.e., matching or adaptability condition). However, the effects were not uniform across the dimensions of performance, workload and stress.

For perceived workload, Global Workload changes as a function of task demand such that when demand is high, Global Workload increases. This was also observed for the stress subscales, with the exception of Temporal Demand. For that scale, the deleterious effect of experiencing adaptive automation that is not well-mapped to task demands was observed only for automation of high reliability. Even in this case, however, Temporal Demand is higher when task demands are higher. These results generally accord with
previous research on task demands and automation (e.g., Warm et al., 2008).

For the scales associated with post-task distress (Tense Arousal, Hedonic Tone, and Confidence and Control), only a block effect was observed. Changes in Tense Arousal tended to follow task demand. By contrast, changes in Hedonic Tone and Confidence and Control were observed to decrease from pre-task state to block 1, after which the scores remained relatively stable. Distress is associated with the core relational theme of overload of processing capacity (Matthews et al., 2002). Hence, the threat detection task apparently depleted cognitive resources across all conditions as a result of time on task, although the effects on the arousal component of Distress changed with task demand.

Two of the scales associated with Task Engagement (Energetic Arousal and Motivation) also showed a decline over time regardless of automation characteristics. However, for Energetic Arousal the decline was most prominent from block 2 (high task demand) to block 3 (low task demand), suggesting that subjective arousal can decline due to high task demand even after that demand has been reduced. Task Engagement is associated with the relational theme of commitment to effort (Matthews et al., 2002). The current data indicate that such commitment wanes over time on task in threat detection, a result that accords with previous research (e.g., Szalma et al., 2004), but that the reliability and adaptability of the automation do not substantially impact the arousal or motivational components of Task Engagement.

With the exception of Task-Irrelevant Cognitive Interference, the scales associated with Worry (Self-Esteem, Self-Focused Attention, and Task-Related Cognitive Interference) indicated a progressive decline over time, regardless of condition. Hence, these cognitive dimensions of stress state may not be related to the characteristics of adaptive automation examined in the present study.

There were two dimensions of cognitive state that were influenced by automation reliability and adaptability. The declines in Concentration (a facet of Task Engagement) associated with a poor adaptive automation schedule occurred only for those who received low reliability automation. With respect to Task Irrelevant Cognitive Interference (a facet of the Worry dimension), the direction of the effect of adaptability on cognitive interference depended on reliability condition. Specifically, those who experienced high reliability automation with an aid that was well-mapped to task demand experienced higher levels of cognitive interference compared to those who received an aid that was not well-mapped to task demand. By contrast, in the low reliability condition those who received a well-mapped aid reported less cognitive interference than those in the less well-mapped condition. It may be that when automation support is reliable and is triggered when task demand is high and deactivated when task demand is relatively low, the automation support frees cognitive resources that participants can allocate to non-task thoughts. However, such processes may be decreased as a function of well-mapped automation when the reliability of that automation is low. Individuals in those conditions may not have had the resource to devote to task-irrelevant cognitions because the low reliability required them to focus more on task related processing.

In general, the results of the present study confirm the multidimensionality of stress state (Matthews et al., 2002), and indicate that the complex, multidimensionality of stress extends to threat detection tasks that utilize adaptive automation. Hence, the pattern of stress and workload is not likely to be constant for all forms of automation. Future research should examine how the workload and stress profiles associated with performing tasks using adaptive automation changes as a function of other characteristics of automation. For instance, the trigger for the automation in this study was change in task demand. Other triggering criteria, such as the operator’s own performance or physiological indicators of workload or task engagement, may show a different pattern of cognitive state. Similarly, the level of automation as well as the different forms of automation (e.g. decision vs. information automation) may each be associated with different profiles of cognitive state. As higher levels of workload and stress reflect the ‘hidden cost’ of performance (Hockey, 1997) and reveal vulnerability to performance failure, identification of these profiles is an important matter for future research. As technology continues to advance, the role of the human will likely continue to be supplemented by automation. A thorough understanding of how this impacts the human in the system is necessary to insure that automation is not implemented at the expense of the operator’s cognitive well-being.

REFERENCES


