Supervisory Control of Multiple Robots: Effects of Imperfect Automation and Individual Differences

Jessie Y. C. Chen and Michael J. Barnes, U.S. Army Research Laboratory, Orlando, Florida

Objective: A military multitasking environment was simulated to examine the effects of an intelligent agent, RoboLeader, on the performance of robotics operators.

Background: The participants’ task was to manage a team of ground robots with the assistance of RoboLeader, an intelligent agent capable of coordinating the robots and changing their routes on the basis of battlefield developments.

Method: In the first experiment, RoboLeader was perfectly reliable; in the second experiment, RoboLeader’s recommendations were manipulated to be either false-alarm prone or miss prone, with a reliability level of either 60% or 90%. The visual density of the targeting environment was manipulated by the presence or absence of friendly soldiers.

Results: RoboLeader, when perfectly reliable, was helpful in reducing the overall mission times. The type of RoboLeader imperfection (false-alarm vs. miss prone) affected operators’ performance of tasks involving visual scanning (target detection, route editing, and situation awareness). There was a consistent effect of visual density (clutter of the visual scene) for multiple performance measures. Participants’ attentional control and video gaming experience affected their overall multitasking performance. In both experiments, participants with greater spatial ability consistently outperformed their low-spatial-ability counterparts in tasks that required effective visual scanning.

Conclusion: Intelligent agents, such as RoboLeader, can benefit the overall human-robot teaming performance. However, the effects of type of agent unreliability, tasking requirements, and individual differences have complex effects on human-agent interaction.

Application: The current results will facilitate the implementation of robots in military settings and will provide useful data to designs of systems for multirobot control.

Keywords: human-robot interaction, supervisory control, intelligent agent, military, imperfect automation, individual differences, spatial ability, attentional control, gaming experience

INTRODUCTION

Robots are being used more frequently in military operations, and the tasks they are being used for are evolving in complexity. In the future battlefield, soldiers may be given multiple tasks to perform concurrently, such as navigating a robot while conducting surveillance, maintaining local security and situation awareness (SA), and communicating with fellow team members. The possibility of a robotic battlefield creates a number of human factors as well as ethical issues related to nonhuman intelligence conducting combat missions (Barnes & Evans, 2010; Singer, 2010). A potential issue is that the proliferation of intelligent systems could easily overwhelm the human operators’ ability to adequately supervise these systems (Chen, Durlach, Sloan, & Bowens, 2008; Schurr, 2007). As the size of the robot team increases, human operators may fail to maintain adequate SA when their attention is constantly switching between the robots; cognitive resources may also be overwhelmed by the numerous intervention requests from the robots (Lewis et al., 2010). Research shows that autonomous cooperation between robots can aid the performance of the human operators and enhance the overall human–robot team performance (Lewis et al., 2010; Schurr, 2007). However, human operators’ involvement in mixed-initiative teams will still be required for the foreseeable future, especially in situations involving critical decision making. Human operators’ decision making may be influenced by “implicit goals” that the robots are not aware of (i.e., are not programmed into the behaviors of the robots; Linegang et al., 2006) and by real-time developments on the battlefield that may require the human operator to change plans for individual robots or the entire robotic team. Effective communication between the human operator and robots, therefore, becomes critical in ensuring mission success.
Past research has demonstrated the effectiveness of a robot proxy to enhance shared understanding between the human operator and the robot in an exploration task (Stubbs, Wettergreen, & Nourbakhsh, 2008). Other research and development efforts more related to supervisory control of military robotics include the U.S. Army’s Playbook program and the U.S. Air Force’s Vigilant Spirit Control Station for management of multiple unmanned aerial systems (UASes; Fern et al., 2011; Miller & Parasuraman, 2007).

**Current Study**

To achieve a better balance of enhancing autonomy and capability while simplifying human-robot interaction, a robotic surrogate for the human operator, RoboLeader, was developed under the U.S. Army Research Laboratory’s (ARL) Director’s Research Initiative Program (Snyder, Qu, Chen, & Barnes, 2010). RoboLeader is an intelligent agent that interprets an operator’s intent and issues detailed command signals to a team of robots of lower capabilities. Instead of directly managing each individual robot, the human operator deals only with a single entity, RoboLeader. The operator can, therefore, better focus on other tasks requiring his or her attention.

The present research focuses on issues related to human-agent teams: the dynamics of using an agent as a subordinate supervisor to aid in multirobot control as a function of mission task load and agent reliability as well as individual differences among human operators. Specifically, we were interested in understanding the effects of using an intelligent agent as a team member in highly demanding multitasking environments when its suggestions were less than perfect and in understanding how individual differences interacted with agent unreliability. To date, there has been minimal research on supervisory control investigating all these factors in one study.

We examined the effectiveness of RoboLeader for enhancing the overall human-robot teaming performance in two experiments. In Experiment 1, we compared the operators’ target detection performance in the four-robot and eight-robot conditions, with or without RoboLeader. In Experiment 2, we investigated the effects of various reliability levels for RoboLeader on operators’ multitasking performance; the reliability of RoboLeader’s recommendations was manipulated to be either false-alarm prone (FAP) or miss prone (MP), with a reliability level of either 60% or 90%. In both experiments, the effects of individual differences in spatial ability, attentional control, and gaming experiences were examined. In the rest of this section, we briefly review the main theoretical issues investigated in these experiments. For a detailed review on human performance issues related to supervisory control of multiple robots, see Chen, Barnes, and Harper-Sciarini (2011).

**Imperfect Automation**

Since it is unlikely for any automated systems to achieve 100% reliability at all times, the effects of unreliable automation on human operator performance need to be better understood before these intelligent systems can be implemented. The effects of imperfect automation are examined by Meyer (2001, 2004), who suggests that FAP and MP alerts may affect the use of an automated system in different ways. High false alarm (FA) rates were seen to reduce the operator’s response to alerts (i.e., compliance), and high miss rates reduced the operator’s reliance on automated systems. Similar results were reported in Dixon and Wickens (2006). They found that the operator’s performance of the automated task degraded when the automation was FAP because of the operator’s reduced compliance with the automated system; when the miss rate was high, the operator’s performance of the concurrent task was affected more than the automated task because the operator allocated more visual attention to monitor the automated task.

McCarley (2007) showed that FAP automation hurt performance more on the automated task than did MP automation (“cry wolf” effect) and hurt performance at least as much as MP automation on the concurrent task (pp. 570–571). Finally, Wickens and Dixon (2007) demonstrated that when the reliability level is below approximately 70%, operators will often ignore the alerts. In their meta-analytic study, Wickens and Dixon found that “a reliability of 0.70 was the ‘crossover point’ below which unreliable
Supervisory Control of Multiple Robots

automation was worse than no automation at all” (p. 201). In our first experiment, which serves as a baseline investigation of the utility of RoboLeader, a perfect reliability level was implemented. In the second experiment, we manipulated both RoboLeader’s error type and reliability level to systematically examine the effects of automation imperfections on the overall human-robot team performance.

Individual Differences

In the current study, we also sought to evaluate whether individual differences in spatial ability, attentional control, and video gaming experience might affect the operator’s performance. Spatial ability (SpA) has been found to be a significant factor in certain visual display domains (Stanney & Salvendy, 1995), multitasking involving flight asset monitoring and management (Morgan et al., 2011), virtual environment navigation (Chen, Czerwinski, & Macredie, 2000), target search task (Chen, 2010; Chen et al., 2008; Chen & Joyner, 2009; Chen & Terrence, 2008, 2009), and robotics task performance (Cassenti, Kelley, Swoboda, & Patton, 2009; Lathan & Tracey, 2002). U.S. Air Force scientists (Chappelle, McMillan, Novy, & McDonald, 2010; Chappelle, Novy, Randall, & McDonald, 2010) interviewed 53 subject matter experts about abilities that were critical to effective performance of UAS control tasks in terms of piloting and sensor operations; SpA was identified as an important factor for both tasks. Our previous research showed that individuals with greater SpA exhibited more effective visual scanning and target detection performance (Chen, 2010; Chen et al., 2008; Chen & Joyner, 2009; Chen & Terrence, 2008, 2009).

In addition to spatial ability, the relationship between attentional control and multitasking performance was also examined. Attentional control is defined as one’s ability to focus and shift attention in a flexible manner (Derryberry & Reed, 2002). Several studies have shown that there are individual differences in multitasking performance, and some people are less prone to performance degradation during multitasking conditions (Rubinstein, Meyer, & Evans, 2001; Schumacher et al., 2001). There is evidence that people with better attention control can allocate their attention more flexibly and effectively, and attention-switching flexibility can predict performance of such diverse tasks as flight training and driving (Bleckley, Durso, Crutchfield, Engle, & Khanna, 2003; Derryberry & Reed, 2002; Feldman Barrett, Tugade, & Engle, 2004; Kahneman, Ben-Ishai, & Lotan, 1973). According to a recent U.S. Air Force survey of subject matter experts on the performance of operators of UASes (Chapelle, McMillan et al., 2010), attentional control is one of the most important abilities that would affect an operator’s performance, since the robotics control task is inherently multitasking (e.g., sensor manipulation, tracking, communication).

According to Feldman Barrett et al. (2004), those with lower attentional control tend to take the “cognitive miser” approach (i.e., conserving the amount of cognitive resources) when dealing with complex information processing to reduce the attentional control requirements. When dealing with automated systems, therefore, it is likely that operators with different levels of attention switching abilities may react differently to unreliable automated systems. In other words, operators’ behaviors of compliance with, and reliance on, automation may be altered by their ability to effectively switch their attention among the systems. For example, the automation-induced complacency effect repeatedly demonstrated in previous research (Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Parasuraman & Manzey, 2010; Thomas & Wickens, 2004; Young & Stanton, 2007) may be more severe for individuals with poor attentional control compared with those with better attentional control. This phenomenon has been demonstrated in Chen and Terrence (2009) and was further tested in the current experiment, which employed a simulation environment considerably different from that of Chen and Terrence’s.

Finally, the current study sought to examine the relationship between participants’ video gaming experience and their task performance as well as SA of the mission environment. According to Green and Bavelier (2006) and Hubert-Wallander, Green, and Bavelier (2010), experienced action video game players, compared with infrequent gamers and nongamers, were found to
perform significantly better on tasks that required visuospatial selective attention, multiple object tracking, rapid process of visual information and imagery, and flexibility in attention allocation. Hambrick, Oswald, Darowski, Rench, and Brou (2010) also demonstrated the relationship between video game experience and multitasking performance. Therefore, we expected frequent gamers would outperform infrequent gamers in our visually demanding task environment in terms of target and event detection and monitoring of the mission environment.

**EXPERIMENT 1**

In Experiment 1, a military reconnaissance mission environment was simulated to examine the effectiveness of RoboLeader, an intelligent agent capable of coordinating a team of ground robots and changing their routes on the basis of developments in the mission environment. The participants’ task was to manage either four or eight robots with or without the assistance of RoboLeader while searching for hostile targets via streaming video from the robots. The experiment is a mixed-model design, with RoboLeader (with or without RoboLeader [baseline]) as the between-subject variable and the number of robots used in the scenario (four vs. eight) as the within-subject variable.

**Method**

**Participants.** A total of 30 individuals from the Orlando, Florida, area (17 males and 13 females, mean age 24.7) participated in the study. Participants were compensated $15 per hour for their participation in the experiment.

**Simulators.** A modified version of the Mixed Initiative Experimental (MIX) Testbed was used as the simulator for this experiment (Barber, Davis, Nicholson, Finkelstein, & Chen, 2008). The Operator Control Unit (OCU) of the MIX Testbed (Figure 1) was modeled after the Tactical Control Unit developed under the ARL Robotics Collaborative Technology Alliance. RoboLeader has the capability of collecting information from subordinate robots with limited autonomy (e.g., collision avoidance and self-guidance to reach target locations), making tactical decisions, and coordinating the robots by issuing commands, waypoints, or motion trajectories (Snyder et al., 2010). See the Procedure
section for more details about the user interface of RoboLeader.

**Surveys and tests.** A demographics questionnaire was administered at the beginning of the training session. An Ishihara color vision test (with nine test plates) was administered via PowerPoint presentation. The RoboLeader user interface employed several colors to display the plans for the robots, and normal color vision was required to effectively interact with the system. A questionnaire on attentional control (Derryberry & Reed, 2002) was used to evaluate participants’ perceived attentional control (PAC). The Attentional Control survey consists of 21 items and measures attention focus and shifting. The scale has been shown to have good internal reliability ($\alpha = .88$). The Cube Comparison Test (Ekstrom, French, & Harman, 1976) and the Spatial Orientation Test (Gugerty & Brooks, 2004) were used to assess participants’ spatial ability. Participants’ perceived workload was evaluated with use of the computer-based version of the NASA Task Load Index questionnaire (NASA-TLX; Hart & Staveland, 1988).

**Procedure.** Participants were randomly assigned to the RoboLeader group or the baseline (no RoboLeader) group before their sessions started. Before the training session, they completed the preliminary tests (color vision and spatial) and surveys (demographic and PAC). Training, lasting about 1 hr, was self-paced and was delivered by PowerPoint slides showing the elements of the OCU, steps for completing various tasks, several mini exercises for practicing the steps, and exercises for performing the experimental tasks. The participants had to demonstrate that they could recall all the steps for performing the tasks without any help. The experimental session immediately followed the training session and consisted of two scenarios, each lasting approximately 30 min, in which participants used their robotic assets to locate 20 targets (i.e., 10 insurgents carrying weapons and 10 improvised explosive devices [IEDs]) in the remote environment. There were four robots available in one scenario and eight robots in the other scenario. The order of scenarios was counterbalanced across participants.

When each scenario started, the robots began by following preplanned routes, at which time the operator’s task of monitoring the environment and detecting insurgents or IEDs began (by clicking on one of the four video thumbnail views highlighted in the color of its route to enlarge it into the expanded view on the top of the screen). Because of the size of the thumbnails, the participants needed to enlarge them into the expanded view to identify the targets. The robots did not have aided target recognition capability; therefore, the participants had to detect the 10 insurgents (individuals dressed in Middle Eastern attire and carrying weapons) and 10 IEDs by themselves.

For the insurgents, participants used their computer mouse to click on the targets to “lase” them as soon as they were detected. The “lased” insurgents were automatically displayed on the map. For the IEDs, the participants clicked on the IED button on the interface and then marked the locations of the IEDs on the map. There were friendly dismounted soldiers (individuals dressed in military uniform) and civilians (individuals dressed in Middle Eastern attire without weapons) in the simulated environment to increase the visual noise for the target detection tasks. The participants were told that their objective was to finish reconnoitering the area using their robots in the least amount of time possible. Therefore, when replanning a route, the participant and/or RoboLeader must consider both the effectiveness and efficiency of the new route.

In each scenario, there were six “events” that required revisions to a robot’s current plans or route. In some cases, robots had to be rerouted to avoid certain areas (e.g., obstructions and hostile areas); in other cases, robots were rerouted to investigate nearby high-priority areas on the basis of intelligence messages. After an event transpired (indicated by visual and/or auditory alerts), the baseline participants must notice that the event had occurred and then reroute the robot that was affected by the event. For the RoboLeader group, RoboLeader would recommend plan revisions to the operator (by presenting the new route to the operator visually on the map), who could either accept the plans (by clicking on a button) or modify them
as necessary (by dragging the waypoints to the desired locations). Out of these six events, three were “bottom-up” (e.g., unanticipated obstacles detected by the robots that obstructed their navigation; indicated by flashing thumbnails) and three “top-down” (e.g., intelligence that the human operator received from the intelligence network; indicated by auditory alerts as well as appearance of icons on the map). The participants were told to do their best to perform both the rerouting and the target detection tasks instead of focusing on one at the expense of the other.

In each scenario, there were five SA queries, which were triggered on the basis of time progression (e.g., 3 min into the scenario). The SA queries included questions such as “Which areas have the robots searched?” (the participants were instructed to mark the searched areas on a blank map), “Which of your robots is the closest to [area of interest]?” and so on. The OCU screen was blank when an SA query was triggered, and only the SA query and the answer box were displayed on the screen. After each scenario, the participants assessed their workload (NASA-TLX) and then took a 2-min break. The entire experimental session lasted approximately 2 hr.

Dependent measures and data analysis. Dependent measures include number of targets located and identified, the operator’s SA of the mission environment as well as awareness of the status of the individual robots, and perceived workload. Mixed-model ANCOVAs were performed with RoboLeader (with or without RoboLeader) as the between-subject factor and number of robots (four vs. eight) as the within-subject factor. Participants’ SpA (composite score of the two spatial tests) and their Attentional Control survey scores were used as covariates.

Results

Target detection performance. Table 1 lists several measures relating to operator performance and perceived workload. In terms of target detection performance, the analysis revealed that participants detected significantly fewer targets (with insurgents and IEDs combined) when they had eight robots compared with the condition when four robots were available, $F(1, 26) = 25.35, p = .0001, \eta^2_p = .494$. Participants with higher SpA (those with higher composite scores of spatial tests) detected significantly more targets than did those with lower SpA, $F(1, 26) = 8.83, p = .02, \eta^2_p = .254$ (Figure 2). The effects of RoboLeader and attentional control were not statistically significant.

SA. The analysis revealed that participants’ SA was significantly poorer in the eight-robot condition compared with the four-robot condition, $F(1, 26) = 13.31, p = .001, \eta^2_p = .33$ (Figure 3). Frequent video game players had significantly better SA than infrequent gamers in the RoboLeader condition, $F(1, 11) = 5.90, p = .02, \eta^2_p = .35$, but not in the baseline condition.

Perceived workload. The analysis showed that participants experienced significantly higher workload when there were eight robots ($M = 69.3$) versus four robots ($M = 64.3$),

<table>
<thead>
<tr>
<th>Measure</th>
<th>4 Robots</th>
<th>8 Robots</th>
<th>4 Robots</th>
<th>8 Robots</th>
<th>Summary of Main Effects$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targets (% detected)</td>
<td>75.0 (10.9)</td>
<td>61.3 (13.9)</td>
<td>77.0 (14.6)</td>
<td>62.7 (18.9)</td>
<td>4 robots &gt; 8 robots</td>
</tr>
<tr>
<td>SA queries (% correct)</td>
<td>46.3 (17.2)</td>
<td>26.7 (25.8)</td>
<td>48.7 (22.6)</td>
<td>23.8 (20.2)</td>
<td>4 robots &gt; 8 robots</td>
</tr>
<tr>
<td>Workload (NASA Task Load Index)</td>
<td>67.4 (16.3)</td>
<td>71.4 (18.1)</td>
<td>61.3 (14.9)</td>
<td>67.1 (14.7)</td>
<td>8 robots &gt; 4 robots</td>
</tr>
</tbody>
</table>

Note. Standard deviations are presented in parentheses.

$^a$Based on statistical analyses reported in the Results section.
Supervisory Control of Multiple Robots

F(1, 27) = 5.24, p = .023, η²_p = .158 (Figure 4). Participants in the RoboLeader group assessed their workload slightly lower (M = 64.1) than did those in the baseline group (M = 69.4). However, the difference failed to reach statistical significance. Participants with higher PAC rated their workload as significantly lower than did those with lower PAC; similarly, frequent video gamers’ workload assessments were significantly lower than those of infrequent gamers’, p < .05 (Figure 4).

Operator’s interaction with the OCU. Participants’ interaction with the OCU (e.g., clicks on the graphical control interface) was analyzed. Participants’ SpA was significantly correlated with the number of thumbnail clicks (to expand the video to find targets), r = .448, p = .006. Additionally, participants in the RoboLeader group spent significantly less time completing their mission scenarios (not including the time on modifying the plans) than did those in the baseline group (20.7 vs. 23.8 min), F(1, 27) = 7.12, p = .013, η²_p = .209. The longer completion times in the baseline condition suggest that the robots’ routes were less efficient than those in the RoboLeader condition, since both groups were alerted about the events that required route revisions and the time participant spent on rerouting was excluded from the mission completion times.

EXPERIMENT 2

In Experiment 2, the effects of unreliable RoboLeader on operator performance were investigated. The participants’ task, as in Experiment 1, was to manage four robots with the assistance of RoboLeader while searching for hostile targets via streaming video from the robots. The reliability of RoboLeader’s solutions was manipulated to be either FAP or MP, with a reliability level of either 60% or 90%. Furthermore, Experiment 2 simulated a multitasking environment rather than a
dual-tasking environment as in Experiment 1. In addition to the target detection and route revision tasks, the participants had to simultaneously perform a gauge monitoring task and an auditory communication task. Finally, the visual density (VD) of the simulated environment (clutter of the visual scene) was manipulated; there were twice as many entities (virtual human characters) in the high-density environment as in the low-density environment. The experiment is a mixed-model design, with RoboLeader imperfection type (FAP vs. MP) and reliability level (60% vs. 90%) as the between-subject factors and VD (high vs. low) of the simulated environment as the within-subject factor.

Method

Participants. For the second experiment, 40 individuals (23 males and 17 females, mean age 23.8 years) from the Orlando, Florida, area participated. They were compensated $15 per hour for their time.

Apparatus. The simulator in the first experiment was modified so the reliability of RoboLeader was either FAP or MP, at a level of either 60% or 90% (FAP60, FAP90, MP60, and MP90, respectively; see Procedure). The number of robots was reduced to four and a gauge monitoring task and a communication task (a text input area and a submission button) were added to the user interface (Figure 5). More details about these manipulations and changes are presented in the Procedure section.

The same surveys and tests used in Experiment 1 were employed in Experiment 2. A modified version of the Usability and Trust Questionnaire used in Chen and Terrence (2009) assessed participants’ perceived usability of the RoboLeader system as well their trust in the system. The items that measured participants’ trust in the system were modified from the Trust Between People and Automation questionnaire (Jian, Bisantz, & Drury, 2000). The questionnaire consisted of 22 questions on a scale of 1 to 7 and included items such as “The RoboLeader display can be deceptive” and “The RoboLeader system is dependable.”

Procedure. Participants were randomly assigned to the FAP60, FAP90, MP60, or MP90 group (with 10 participants per group) before their sessions started. The procedure of the pre-experimental session (surveys, tests, and training) followed the procedure of Experiment 1 and lasted approximately 1 hr. The type and
reliability level of each participant’s RoboLeader condition (FAP60, FAP90, MP60, or MP90) matched in the training and experimental scenarios. The participants were informed that RoboLeader was either FAP or MP and “fairly but not always reliable” (for the 90% conditions) or “not always reliable” (for the 60% conditions).

The experimental session began immediately after the training session and lasted about 1 hr. Each experimental session had two scenarios (one with high VD and one with low density), both lasting approximately 30 min. During the scenarios, participants used their four robots to locate 10 targets (insurgents carrying weapons) while rerouting their robots around events in the remote environment (described later). In the low-density scenario, there were about 600 civilians throughout the scenario, and in the high-density scenario, there were about 600 civilians and 600 friendly soldiers visible in the environment. The presence of friendly soldiers in the high-density scenario made the target detection task more difficult, as the friendly soldiers all carry weapons. The procedure of the target search task followed that of Experiment 1, and the order of scenarios was counterbalanced across participants.

During the scenarios, as in Experiment 1, several events required revisions to the robots’ routes. RoboLeader and the participants needed to create new routes toward “high-priority areas” while avoiding rerouting the robots through “hostile areas” or areas already traversed. As in Experiment 1, participants were told that their objective was to finish reconnoitering the area using their robots in the least amount of time possible while keeping all route edits as close as possible to the original routes. In the FAP conditions, RoboLeader would provide rerouting recommendations that were not necessary. Participants could check the validity of RoboLeader’s recommendations by reviewing their map. A true event was associated with an icon (a red square for a hostile area and a blue square for a high-priority area; see Figure 5), but FAs were not.

In the FAP60 scenario, following the signal detection theory paradigm (Green & Swets, 1988), RoboLeader provided solutions to all five events that required revisions to robots’ routes, and it also provided solutions on four occasions when no events occurred and no revisions were necessary (i.e., five hits, four FAs, zero misses, and one correct rejection), making a total of 10 events, 6 of which were positive. In the FAP90 scenario, RoboLeader provided solutions to all 5 events that required revisions, and it also provided solutions on one occasion when no events occurred. In the MP scenarios, RoboLeader would fail to provide solutions when some events happened. In the MP60 scenario, 10 true events occurred that required revisions to a robot’s route, although RoboLeader provided solutions for only 2 of them. In the MP90 scenario, 10 true events occurred and RoboLeader provided solutions for 8 of them.

In addition to the tasks described already, the participants simultaneously performed a gauge monitoring task and an auditory communications task. The gauge monitoring task (upper left corner of the OCU under the blue RoboLeader message box) displayed four gauges constantly in motion that entered an upper or lower limit at various prespecified times throughout the scenarios. The participants were required to monitor the gauges and press a “Reset” button when any gauge entered the upper or lower limit to put the gauges back to their normal levels. The auditory communications task presented prerecorded questions at 30-s intervals during the scenarios. Questions included simple military-related reasoning and memory (i.e., call-sign recognition) tests. Participants used a keyboard to enter their responses for the questions into the communications panel on the OCU (adjacent to the gauges; see Figure 5). As in Experiment 1, the participants were told to do their best to perform all the tasks instead of focusing on some at the expense of the others. Each scenario also contained five SA queries, which followed the same format and procedure as in Experiment 1.

A 2-min break was given between the experimental scenarios. Participants assessed their workload using an electronic NASA-TLX immediately after each scenario. Following completion of both scenarios, participants were asked to evaluate the usability of the RoboLeader
system by filling out the Usability and Trust Questionnaire.

Dependent measures and data analysis. Dependent measures include the number of targets located and identified, the number of routes edited, the operators’ SA of the mission environment, concurrent task performance (gauge monitoring and auditory communications), and perceived workload. Mixed-model ANCOVAs were performed with RoboLeader imperfection type (FAP vs. MP) and reliability level (60% [low] vs. 90% [high]) as the between-subject factors and VD (high vs. low) as the within-subject factor. Participants’ composite SpA test scores and their Attentional Control survey scores were used as covariates.

Results

Route editing task performance. Table 2 lists several measures relating to operator performance and perceived workload. In terms of the operator’s route editing (i.e., automated) task performance, the analysis showed that both imperfection type and reliability level of RoboLeader significantly affected the percentage of routes that participants edited, $F(1, 35) = 161.7, p = .0001, \eta^2_p = .82, and F(1, 35) = 7.4, p = .01, \eta^2_p = .18$, respectively (Figure 6). Participants edited more routes in the FAP condition and in the high-reliability (90%) condition. There was a significant difference between those with higher SpA and lower SpA, $F(1, 35) = 4.1, p = .045, \eta^2_p = .104$ (Figure 7). Further analysis revealed that high-SpA participants outperformed low-SpA participants to a greater extent in the high-VD condition than in the low-VD condition, $F(1, 37) = 6.09, p = .018, \eta^2_p = .141$.

SA. Participants’ SA (percentage of SA queries answered correctly) was significantly better in the MP condition than in the FAP condition, $F(1, 35) = 8.5, p = .003, \eta^2_p = .20$ (Figure 8). Frequent video game players had slightly better SA than did infrequent gamers; however, the difference failed to reach statistical significance, $F(1, 38) = 3.33, p = .076$.

Communication task performance. There were no main effects associated with the communication task performance. There was a significant interaction between VD and participants’ PAC, $F(1, 35) = 5.4, p = .026, \eta^2_p = .134$.

Gauge monitoring performance. There was a significant interaction between VD and reliability level of RoboLeader, $F(1, 35) = 4.3, p = .047, \eta^2_p = .11$. Higher-PAC participants responded faster than those with lower PAC, $F(1, 35) = 84.7, p = .001, \eta^2_p = .31$ (Figure 9). The difference between high- and low-PAC participants was especially pronounced in low-reliability conditions than in high-reliability conditions. Frequent video game players also outperformed infrequent gamers, $F(1, 38) = 8.19, p = .009, \eta^2_p = .18$.

Individual differences in multitasking performance. To further test the effects of individual differences in operators’ overall multitasking performance, multivariate ANOVAs were performed on all the task performance data. There was a main effect of PAC, $F(5, 33) = 3.60, p = .011, \eta^2_p = .35$. Further analysis revealed that high-PAC participants outperformed low-PAC participants in the MP condition, $F(5, 14) = 4.50, p = .012, \eta^2_p = .62$. In the FAP condition, however, the difference failed to reach statistical significance. There was also a main effect of participants’ video gaming experience, $F(5, 34) = 2.42, p = .04, \eta^2_p = .26$. There failed to be an overall effect of operator’s SpA on task performance.
### TABLE 2: Experiment 2: Mean Operator Task Performance and Workload Assessments

<table>
<thead>
<tr>
<th>Measure</th>
<th>60%</th>
<th>90%</th>
<th>60%</th>
<th>90%</th>
<th>Summary of Main Effects¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Alarm Prone (FAP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Route editing (% correct)</td>
<td>95.6 (7.7)</td>
<td>97.8 (4.6)</td>
<td>95.0 (11.2)</td>
<td>98.3 (5.4)</td>
<td>65.0 (13.5)</td>
</tr>
<tr>
<td>Target detection (% correct)</td>
<td>80.0 (11.5)</td>
<td>96.0 (5.2)</td>
<td>72.0 (31.2)</td>
<td>86.0 (14.3)</td>
<td>53.0 (14.2)</td>
</tr>
<tr>
<td>Situation awareness queries (% correct)</td>
<td>16.0 (12.6)</td>
<td>18.0 (19.9)</td>
<td>26.0 (9.6)</td>
<td>20.0 (23.1)</td>
<td>32.0 (19.3)</td>
</tr>
<tr>
<td>Gauge monitoring, reaction time (seconds)</td>
<td>3.91 (1.49)</td>
<td>4.27 (1.87)</td>
<td>4.48 (2.22)</td>
<td>4.16 (3.14)</td>
<td>3.99 (2.27)</td>
</tr>
<tr>
<td>Communication (% correct)</td>
<td>87.7 (4.7)</td>
<td>80.7 (6.0)</td>
<td>82.0 (8.5)</td>
<td>86.7 (8.3)</td>
<td>82.3 (6.5)</td>
</tr>
<tr>
<td>Workload (NASA Task Load Index)</td>
<td>79.4 (10.5)</td>
<td>70.7 (7.3)</td>
<td>67.1 (13.6)</td>
<td>60.0 (17.4)</td>
<td>79.8 (14.4)</td>
</tr>
<tr>
<td>Miss Prone (MP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Route editing (% correct)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67.0 (11.6)</td>
</tr>
<tr>
<td>Target detection (% correct)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>57.0 (15.7)</td>
</tr>
<tr>
<td>Situation awareness queries (% correct)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40.0 (16.3)</td>
</tr>
<tr>
<td>Gauge monitoring, reaction time (seconds)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.13 (1.53)</td>
</tr>
<tr>
<td>Communication (% correct)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82.3 (7.4)</td>
</tr>
<tr>
<td>Workload (NASA Task Load Index)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>75.1 (14.7)</td>
</tr>
</tbody>
</table>

Note. Standard deviations are presented in parentheses. VD = visual density.

¹Based on statistical analyses reported in the Results section.
Perceived workload. The analysis showed that both VD of the target environment and the reliability level of RoboLeader contributed significantly to the participants’ perceived workload, $F(1, 36) = 7.7, p = .009, \eta^2_p = .18$, and $F(1, 36) = 4.8, p = .036, \eta^2_p = .12$, respectively (Figure 10). Participants experienced higher workload in the high-VD condition as well as when the reliability level was lower.

Operator’s interaction with the OCU. Participants made significantly more thumbnail clicks in the low-VD condition than in the high-VD condition, $F(1, 35) = 6.5, p = .006, \eta^2_p = .16$. Participants’ SpA was also significantly correlated with the number of thumbnail clicks, $r = .33, p = .019$. There was no significant difference between higher- and lower-PAC participants in the aggregate scores of the trust questionnaire.

**GENERAL DISCUSSION**

We investigated, in two simulation experiments, the effectiveness of human operators’ performance of military tasks when interacting
with an intelligent agent, RoboLeader, with perfect or less-than-perfect reliability. In both experiments, participants’ main task was to supervise a team of ground robots while detecting hostile targets via streaming video from the robots. Whereas the first experiment involved only the robotic route revision task and the target detection task, the second experiment simulated a multitasking environment in which the participants had to simultaneously perform an additional gauge monitoring task and an auditory communication task.

In the first experiment, although we did not find significant differences in target detection between the two groups, the participants in the RoboLeader group completed their missions in significantly less time than did those in the baseline group. On average, RoboLeader saved the participants approximately 3 min for missions lasting 20 or more minutes (not including time on plan revisions). This suggests that the RoboLeader’s plans were more efficient than the operators’. The results of Experiment 2 showed that the type of RoboLeader imperfection and its reliability level both significantly affected the participants’ route editing performance. As expected, participants edited more routes in the FAP condition and when RoboLeader was highly reliable. Participants’ spatial ability and attentional control both played a role in their performance of route editing, with the performance differences especially pronounced in the MP and low-reliability conditions.

In both experiments, one of participants’ tasks was to find targets via streaming video from the robots. In Experiment 1, participants detected significantly fewer targets when there were eight robots compared with the four-robot condition, indicating less efficiency with more resources or assets. This result is consistent with the finding of a recent study by Squire and Parasuraman (2010), in which the participants performed worse (i.e., won fewer games) with eight robots versus four robots. The results of Experiment 2 showed that participants detected fewer targets when the RoboLeader was MP or when target VD was high. This finding is consistent with that of Dixon and Wickens (2006) that concurrent tasks (compared with automated tasks) are affected more by MP systems than by FAP systems because the operator allocates more visual attention to monitor the automated task when the automation is MP. However, the effectiveness of this elevated visual attention to the automated task was modulated by one’s SpA, as we explain later.

From the performance data, it does not appear that the difference in the numbers of routes that participants had to manually edit had any noticeable effects on their performance of the concurrent tasks. For example, for the target detection tasks, there was not a significant difference between the MP60 and MP90 conditions, although participants were required to change eight routes on their own in the MP60 condition and only two routes in the MP90 condition. Similarly, in Experiment 1, when RoboLeader was perfect and the number of events was the same, there was no difference between RoboLeader and baseline-manual conditions although the manual conditions required six edits and the RoboLeader required none. Taken together, these results suggest that the number of manual route edits required was not the crucial factor. It is, therefore, more likely that the continuous visual monitoring requirement associated with the MP conditions (and factors such as VD) contributed more to the performance decrements observed for the MP condition (although the MP group did have better SA, which is discussed later).

Figure 10. Perceived workload.
In both experiments, those participants with higher SpA detected more targets (especially in high-VD conditions in Experiment 2) and made more thumbnail clicks than did those with lower SpA. These results are consistent with previous findings (Chen, 2010; Chen et al., 2008; Chen & Joyner, 2009; Chen & Terrence, 2008, 2009) that individuals with higher SpA tend to exhibit more effective scanning performance and, therefore, are able to detect more targets than do those with lower SpA, especially when visual processing load is heavy (high-VD conditions in Experiment 2). The fact that high-SpA participants edited more routes than did low-SpA participants in Experiment 2 also suggests more effective scanning by the high-SpA individuals. In the MP60 condition, which required the most visual monitoring by the participants, high-SpA participants were able to edit almost 20% more of the routes than were their lower-SpA counterparts.

These findings support the recommendations by Lathan and Tracey (2002) and two recent U.S. Air Force studies (Chappelle, McMillan, et al., 2010; Chappelle, Novy, et al., 2010) that military missions can benefit from selecting personnel with higher SpA to operate robotic devices. Training interventions that could enhance the spatial interpretations required to perform a mission task might also be of benefit (Rodes, Brooks, & Gugerty, 2005).

In terms of awareness of the overall mission environments, participants’ SA was significantly better when there were four robots than when there were eight robots to keep track of. This result is not surprising and is consistent with previous findings (Crandall, Goodrich, Olsen, & Nielsen, 2005; Squire & Parasuraman, 2010; Trouvain & Wolf, 2002). Participants’ SA was also better in the MP condition than in the FAP condition. This suggests that the participants scanned the map more frequently in the MP condition than in the FAP condition. Again, this is consistent with Dixon and Wickens’s (2006) finding that MP systems drew operators’ visual attention away from the concurrent tasks to focus more on the automated tasking environment. Interestingly, participants’ self-assessed trust in the RoboLeader system did not differ significantly between the MP and FAP groups.

Although we did not collect eye movement data, we can infer some of the attentional processes and strategies of the participants from the performance data. In their effort to maintain optimal performance across the tasks, the participants appeared to allocate more attentional resources to the MP tasking environment than to the FAP environment, as the elevated SA performance for the MP condition suggested. However, this biased attention allocation came with a price—as the degraded target detection performance in MP conditions indicated.

There was some evidence that frequent video gamers tended to have somewhat better SA than did infrequent gamers, although the results were not always consistent (e.g., only in the RoboLeader condition but not in the baseline condition in Experiment 1; difference not reaching statistical significance in Experiment 2). In another RoboLeader study employing different dynamic replanning tasks, it was observed that frequent gamers had significantly better SA of the tasking environment than did infrequent gamers (Chen, Barnes, Quinn, & Plew, 2011). Given findings reported in Green and Bavelier (2006) that frequent gamers tend to have better visual short-term memory, it is not surprising to find them exhibiting better SA of the tasking environments. Additionally, Cummings, Clare, and Hart (2010) found that frequent gamers collaborated more with an automated UAS replanning system (higher degree of consent) than did infrequent gamers. Their finding suggests that the frequent gamers in our Experiment 1 probably worked better with the RoboLeader system than did the infrequent gamers, and this more effective collaboration might have contributed to the frequent gamers’ better SA, especially when RoboLeader was reliable.

Participants experienced significantly higher workload when there were eight robots compared with the four-robot condition, but the presence of RoboLeader did not seem to have a significant effect on their perceived workload. Participants’ workload assessments were also higher when the density of the visual environment was higher or when RoboLeader’s reliability level was lower.

Participants’ PAC was found to have a significant effect on their multitasking performance,
especially the execution of secondary tasks (communication and gauge monitoring). This finding is consistent with those of Chen and Joyner (2009) that participants performed at a similar level on the primary tasks (gunnery and robotics), but those with higher PAC performed better on the secondary communication task than did those with lower PAC. These results suggest that participants with higher PAC were more able to allocate their attentional resources effectively in the multitasking environment than those with lower PAC, especially when tasking environments became more challenging (e.g., low-reliability conditions in Experiment 2). When the automation had low reliability, low-PAC individuals did not appear to be able to allocate as much attention to parts of the tasking environment (e.g., gauges) as did high-PAC individuals (Figure 9). It was also found that participants with higher PAC consistently performed better in the MP condition across different tasks than those with lower PAC. This is consistent with Chen and Terrence’s (2009) finding that MP automated systems tended to be more detrimental to lower PAC individuals than to higher PAC individuals.

An interesting difference between the current results and those of Chen and Terrence (2009) was that in the current study, participants with higher PAC did not exhibit as much undertrust (i.e., disuse) of the FAP system as those high-PAC participants did in the Chen and Terrence study. In the current study, high-PAC participants performed at similar levels as low-PAC participants in the FAP conditions but outperformed low-PAC individuals in the MP conditions. The discrepancy between these results and those of Chen and Terrence may be attributable to the different “costs” of scanning in the two simulated environments. In the Chen and Terrence study, the gunner station and the robotics OCU were displayed on two separate monitors, whereas in the current study, all tasks were performed on the same monitor. Compared with the current study, the cost of scanning in the Chen and Terrence study was greater, and those of higher PAC clearly demonstrated reduced compliance with the FAP automated system. In the current study, high-PAC participants did not show this decrement, likely because of the relative ease of verifying the RoboLeader recommendations on the map by checking the icons.

It is interesting to note that while it was considerably easier to verify the validity of the alerts in the current study, participants with low PAC performed more poorly in the MP conditions than those with high PAC, just as the results of the Chen and Terrence (2009) study showed. A likely reason for this phenomenon is that MP scenarios required continuous scanning of the map to find new icons. This requirement made the task similar to a “change detection” task, although performed in a multitasking environment. The current results suggest that low-PAC individuals cannot detect changes as effectively as their high-PAC counterparts, likely because of their poorer attentional management abilities. The way the low-PAC participants interacted with the automated system in the current experiment was consistent with the “cognitive miser” phenomenon described in Feldman Barrett et al. (2004). The phenomenon states that low-PAC individuals, because of their limited attentional resources, tend to reduce their information processing demands by simplifying their task(s) (e.g., relying on automation to help them with their tasks). Depending on the context, this oversimplification (e.g., overreliance on automation) may have very undesirable consequences (e.g., MP condition) when the aids fail to provide anticipated assistance.

On the other hand, notably, there were no observable trade-off strategies adopted by the low-PAC participants (their performance across the different tasks was uniformly low in the MP condition). In other words, the low-PAC participants were not found to conserve cognitive resources on one task to perform better on the other tasks. This performance decrement, therefore, seems to reflect something more related to attentional abilities than to deliberate strategies. Future research should investigate factors contributing to this deficiency by low-PAC individuals (e.g., working memory capacity and cognitive flexibility; Bühner, König, Pick, & Krumm, 2006; Feldman Barrett et al., 2004; Youmans, Figueroa, & Kramarova, 2011) and possible training or interfaces design strategies to mitigate them.
Participants’ video gaming experience had a significant impact on their overall multitasking performance. These results are consistent with previous findings (Green & Bavelier, 2006) suggesting that video game play is associated with greater visual short-term memory and faster information processing, which in turn may have contributed to game-playing participants’ superior multitasking performance in the current study. These results are consistent with the findings of one recent U.S. Air Force study (McKinley, McIntire, & Funke, 2010) that frequent video gamers outperformed infrequent gamers on robotics (UAS) tasks and, in some cases, performed as well as experienced pilots. These results also support the conclusion of an U.S. Air Force study (Triplett, 2008) based on interviews of UAS pilots that gamers’ superior visual information-processing skills may be able to translate into superior robotics management performance.

CONCLUSION

In the current study, we investigated the effects of an intelligent agent, RoboLeader, on human operators’ performance of supervising multiple robots to complete military reconnaissance missions in a dual-task and a multitasking environment. Overall, it appears that RoboLeader, when perfectly reliable, was effective in reducing the operators’ mission times in target search tasks, although significant benefits of RoboLeader on the operators’ concurrent task performance and workload were not observed.

Results of Experiment 2 show that the type of RoboLeader imperfection affected operator’s performance of tasks involving visual scanning (target detection, route editing, and SA). Furthermore, there was a consistent effect of VD for multiple performance measures. Participants’ self-assessed attentional control and video gaming experience was found to affect their overall multitasking performance. Across experiments, participants with higher SpA consistently outperformed those with lower SpA in tasks that required the most visual scanning (e.g., target detection and thumbnail clicks), regardless of the experimental conditions. Future research should investigate training interventions (e.g., attention management) and/or user interface designs (e.g., multimodal cueing displays) that can mitigate performance shortfalls of those with lower SpA and attentional control (Chen, Barnes, & Harper-Sciarini, 2011; Chen, Haas, & Barnes, 2007; Dux et al., 2009).

ACKNOWLEDGMENTS

This research was supported by U.S. Army Research Laboratory’s Director’s Research Initiative program and SOURCE Army Technology Objective. The authors wish to thank Zhihua Qu, Mark Snyder, Caitlin Kenny, Daniel Barber, David Adams, and Michelle Harper-Sciarini for their contribution to this project.

KEY POINTS

- Intelligent agents, such as RoboLeader, can benefit the overall human-robot teaming performance. However, the effects of unreliable automation, tasking requirements, and individual differences are complex, and their interactions with one another have different ways to affect human-automation interaction in multitasking environments.
- The type of RoboLeader imperfection (false alarm vs. miss prone) affected operator’s performance of tasks involving visual scanning (target detection, route editing, and situation awareness). The effectiveness of participants’ elevated visual attention to the miss-prone environment was modulated by individual differences in spatial ability and attentional control.
- Participants’ attentional control and video gaming experience affected their overall multitasking performance. The positive effects of higher attentional control were most evident when the automated system (RoboLeader) was miss prone.
- In both experiments, participants with higher spatial ability consistently outperformed their low-spatial-ability counterparts in tasks that required effective visual scanning: They scanned the environment faster and detected more targets. The performance differences were especially pronounced when visual processing load was heavy (high visual density and low system reliability).

REFERENCES


Rodes, W., Brooks, J., & Gugerty, L. (2005, May). Using verbal protocols analysis and cognitive modeling to understand strategies used for cardinal direction judgments. Poster presented at the Human Factors of UAVs Workshop, Mesa, AZ.


Jessie Y. C. Chen received her BA in linguistics from National Tsing-Hua University in Taiwan in 1987, her MA in communication studies from the University of Michigan at Ann Arbor in 1989, and her PhD in applied experimental and human factors psychology from University of Central Florida in 2000. She is a research psychologist with U.S. Army Research Laboratory, Human Research and Engineering Directorate, at the field element in Orlando, Florida. Her research interests include human-robot interaction, human-vehicle interaction, supervisory control, and individual differences.

Michael J. Barnes was educated at University of Arizona and New Mexico State University (MA, experimental psychology, 1975). Currently, he works for U.S. Army Research Laboratory (ARL) where he is the ARL lead for the human-robot interaction (HRI) portion of the Army Technology Objective: Safe Operations of Unmanned Reconnaissance in Complex Environments. He currently serves as member of a NATO committee for the HRI of autonomous systems and has coauthored more than 70 articles on the human factors of military systems. His previous human factors positions were as a researcher with the U.S. Navy and as a manager with General Electric. In his present position, he is the ARL chief of the Fort Huachuca, Arizona, field element, where he supports cognitive research in intelligence analysis, unmanned aerial systems, and visualization tools for risk management.

*Date received: May 25, 2011*

*Date accepted: December 19, 2011*