Weapons Director Intelligent-Agent-Assist Task: Procedure and Findings for a Validation Study

Submitted to the 6th International Command and Control Research Technology Symposium, June 2001, Annapolis Maryland
Paper Track: C2 Experimentation

Scott Chaiken

Air Force Research Laboratory,
AFRL/HEAI, Brooks AFB, TX 78235

Linda R. Elliott & Mathieu Dalrymple

Veridian Engineering
2504 Gillingham Drive, Suite 201
Brooks AFB, TX 78235

Michael D. Coovert, Dawn Riddle, Thomas R. Gordon, Kimberly A. Hoffman, Donald E. Miles, and Thomas V. King

Institute of Human Performance, Decision Making, and Cybernetics
University of South Florida, Dept of Psychology, BEH 339
Tampa, FL 33620-8200

Samuel G. Schiflett

Air Force Research Laboratory
AFRL/HEAI, Brooks AFB, TX 78235

Abstract

The Airborne Warning and Control System (AWACS) is a core command and control (C2) function in which sensors, shooters, and refuelers are managed by “Weapons Directors” (WDs) in an airborne radar and communications command post. Improving the quality of WD training can have profound effects on mission outcome. A basic technology capable of doing this is “intelligent-agent” technology, which allows more frequent practice, via simulated players, and embedded decision aids that display reasonable task options online. We report initial empirical work with an embedded-agent simulation based on the AWACS, namely the 21st Century Systems, Inc. Weapon’s Director Intelligent-Agent Assist platform. Using this platform, we observed how 38 WDs performed during two high-workload missions. One mission was played with a decision aid that recommended target pairings and refuelings, while the other was not. Our sample benefited from the decision aid, but the more experienced WDs benefited the most (counter to our expectations). We discuss the results in terms of interface challenges decision aids will face in high workload environments. This report extends the initial report for this system (Elliott, Chaiken, Dalrymple, Petrov, Stoyen, 2000).

Introduction[*]

Combining computer-simulated enemy and friendly combatants within training simulations has the potential of dramatically reducing training cost by allowing realistic team practice without a full complement of players. This would allow frequent and asynchronous practice among future mission participants. Other direct applications of the same technology include embedded training coaches (i.e., decision aids), campaign simulation, job-centered workload measurement, and scientific control for measuring command and control team performance when faced with realistic (but standardized) adversaries. Given this list of potential boons and a computer science that is rapidly becoming capable of delivering intelligent agents, it would seem that the time is right to implement and research such systems.

A moderate-fidelity Airborne Warning and Control Systems (AWACS) task was the context in which we evaluated a new embedded intelligent-agent technology. The agent technology governed the behavior of synthetic forces that the trainee could interact with (e.g., fight in the case of an enemy, assist in the case of a friendly strike force). In addition, the agent technology could serve as decision support. We trained Weapons Directors (WDs) in the task and allowed them to play the simulation with and without the agent available for support. When available, the WD could accept, reject, or ignore online recommendations as the agent assessed the tactical situation in parallel with the WD. Our focus here is on the decision aid's impact on performance. Also of interest were the WDs' perceptions of the agent technology as applied to the AWACS simulation.

The system we investigated, namely the WD-IAA Platform (Petrov, Stoyen, & Myers, 2000), was a contracted effort. This effort followed on a cognitive-task-analysis phase that described cognitive and functional aspects of the AWACS WD team, with particular focus on team interaction and interdependencies (Schiflett, & Elliott, 2000; Elliott, Schiflett, Hollenbeck & Dalrymple, in press; Fahey, Rowe, Dunlap, & DeBoom, 2000; Klinger, Andriole, Militello, Adelman, Klein, & Gomes, 1993; Coovert, Gordon, Foster, Riddle, Miles, Hoffman, & King, 1999; Coovert and Riddle, 1999; Dalrymple, 1991). The platform was designed utilizing the results of these cognitive task analyses and the specific guidance of subject matter experts. In Elliot et. al. (2000), we gave an overview of the general approach, methodology, and potential application areas for agent-technologies within C2 training, highlighting our research plans for this particular task. In this report, we have had more time to fully understand the data collected from our one excursion with this platform, which is targeted for the WD community, as a trainer and, perhaps in the future, as an operational decision aid.

[*] Opinions, interpretations, conclusions, and recommendations are the authors and are not necessarily endorsed by the United States Air Force. This research was supported by work units 1123A117 and 1123A002 through funding by AFOSR and AFRL.

Method

Participants

Our WD participants were from the 552nd Air Combat Wing AWACS WD training squadron located at Tinker AFB. They were first categorized by their experience level. Experience level was determined by the number of flight hours in the E-3 aircraft. Those who had logged more than 400 hours and who had at least 1 year Combat Mission Ready Experience were categorized as experienced. This definition reflects the current policy of the 552nd Training Squadron. By this definition, there were 17 "experienced" and 21 "inexperienced" WDs available for testing. However, the "inexperienced" subjects were still categorized as "Mission Ready," and thus they were not naïve with regard to AWACS WD goals, functions, or taskwork. One subject's data from the Inexperienced group was only partially analyzed (in secondary analyses not involving the assessment of agent benefit) because equipment failure lead to missing data in some conditions. While "inexperienced" or newly-trained WDs were fairly homogeneous in terms of their actual experience, the "experienced" group had a somewhat broader allowable range of experience (i.e., some right on the boundary of "experienced" in terms of the requirement). Nevertheless, 10 of the 17 "experienced" WDs were actually instructors who had much more experience than 400 hours (e.g., 1000 to 3000 hours).

General Procedure

The entire testing session had many phases and lasted about four hours with regular breaks. We administered the task to WDs in groups of four. All WDs sat at a large table in front of a laptop computer with a 16" display, which administered the task. Each WD interacted with their laptop using the mouse. WDs were told not to interact with each other, because each WD had their own air-war scenario to fight (i.e., the laptops were not networked together in team mode, which is another capability of the task). Each WD managed forces that interacted with other agent-managed forces. The agent and human-managed forces are described in a later section.

All subjects first experienced indoctrination and hands-on training. This consisted of a 45-minute briefing based on the 21CSI tutorial for the task, presented by a former AWACS WD instructor, and two low-workload 30-minute training sessions, based on a different geography than our experimental ones. The first 30-minute practice session had the agent available. This session practiced recommendation-accepting aspects of the interface. However, manually implemented orders could still be given and practiced. A recommendation appeared as a pairing line between an asset and an enemy target along with an agent-face icon, positioned midway along that line. A WD could accept specific recommendations by clicking individually on their faces, or accept all recommendations presented by clicking an accept-all button. The second 30-minute practice session used the exact same scenario, but had the agent turned off, so that only the manual mode of issuing orders could be practiced. Manual orders were basically point-and-click assignments of asset to target.

After the orientation/practice, the experimental sessions were given, crossing order (i.e., agent availability in the first or second session), geography (i.e., specific scenario paired with agent availability), and WD experience (experienced or newly trained) in a counterbalanced fashion. Hence the primary conditions are agent-availability, a within-subjects manipulation, and experience level, a between-subjects manipulation.

After the two experimental sessions, our collaborators from the University of South Florida administered another experimental session using a verbal protocol method, along with postsession questionnaires they had specifically designed to assess WD perceptions. During the verbal protocol session, WDs put on headphones and were instructed to "think aloud" while they performed a high-workload scenario using the training geography. Half the WDs had the agent available for this session; half did not. The WD performance data recorded from this procedure has been subjected to different and more detailed kinds of analyses that is reported elsewhere (Coovert, Riddle, Gordon, Miles, Hoffman, King, Elliot, Schiflett, & Chaiken, 2001; Gordon, Coovert, Riddle, Miles, Hoffman, King, Elliott, Schiflett, & Chaiken 2001). Finally, WDs completed the questionnaires containing Likert-type ratings and open-ended queries. WDs were told to base their responses on their experience throughout the entire testing session. These questionnaires constitute important subjective data that is directly relevant to how the WDs view an agent in their workplace.

Scenario Construction

A subject matter expert (Mathieu Dalrymple) constructed three roughly equivalent simulation scenarios. This involved identifying the types of events most likely to affect workload and scripting the scenarios to accommodate these. These events included high-tempo stretches of enemy activity in which many simultaneous intercept decisions had to be made while keeping track of refueling needs. Scenarios had agent-modeled forces in three scenario roles. The three automated agent roles were hostile defensive-counter-air, hostile strike force, and Air Force (AF) strike force. Weapons Directors played Air Force defensive-counter air (AFDCA), the fourth scenario role.

The WDs' primary (AFDCA) mission was to defend friendly air space with a secondary mission of protecting the AF Strike Force (slow moving, vulnerable bombers) as these carried out their primary mission of bombing enemy targets. The two experimental scenarios are given in Appendix A of Petrov, Stoyen, and Myers (2000), so that one could perform the simulations as the WDs did. We will refer to these scenarios using their geographic locations: Taiwan and Cyprus.

Scenarios were used to generate WD "fighter-flow" sheets. These sheets provided a script for when AFDCA assets (primarily fighters) would become available during the scenario run (i.e., when these assets could be launched from base). In addition to the timing issues, the sheets also contained fuel and armament status of assets. Prior to each session WDs studied the appropriate fighter-flow sheet and kept the sheet on-hand for note taking throughout the session. Such sheets are typically used in WD operations as a memory and planning aid.

Finally, an important property of all scenarios was their deterministic nature. Given an interceptor's weapons were in range of a target, the interceptor shot down the target with probability 1.0. While this is (perhaps) not an accurate portrayal of real-world armament, setting the simulation to be deterministic in this fashion allowed performance to be more a function of WD skill and not luck.

Results

Initial Expectations

Our experiment tests the utility of agent recommendations by comparing the performance of newly trained and experienced groups with and without the agent recommendations available. As the agent was designed to capture the decision policies of an “expert” WD, we did not expect the recommendations to benefit expert WD performance. However, the agent was expected to boost the performance of inexperienced WDs. Specifically, we expected inexperienced WDs to: (a) use the agent more frequently (i.e., accept more recommendations), (b) perform as effectively as experienced WDs when the agent was available, but (c) perform at a lower level than the experienced WDs when the agent was not available.

Scenario Outcome Performance Scores

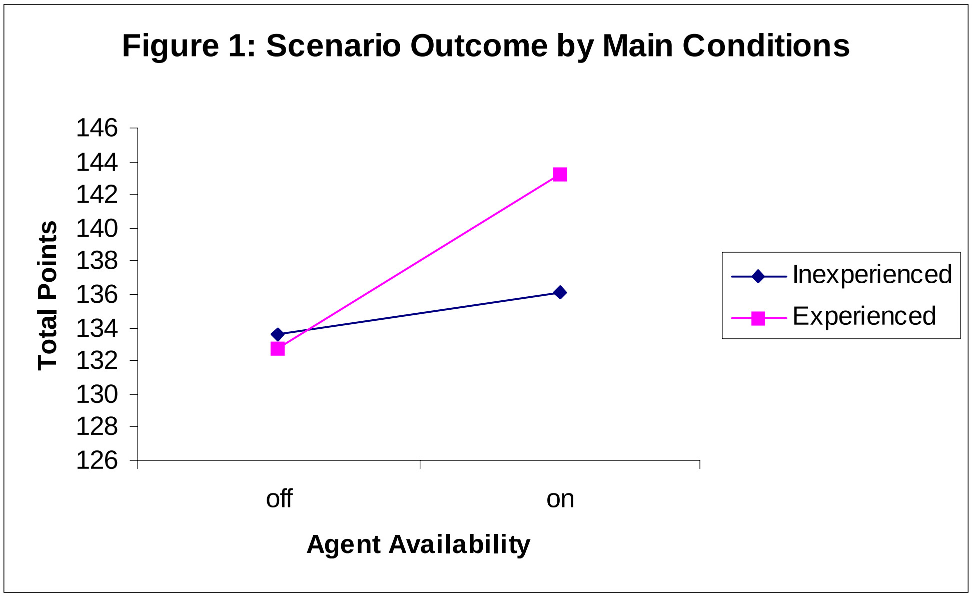
During the 30 minute testing scenarios, a WD’s AFDCA score reflected the number of enemy targets destroyed (points added to score) and the number of AFDCA forces lost (points subtracted from score). The specific kind of targets-destroyed or asset-lost was given point values based on pooled subject-matter-expert’s rankings of the target’s or asset’s (class) value. Additionally, one might consider the score obtained by the AF Strike agent, because the AFDCA’s secondary mission involved protecting the Strike package (so the AF Strike score should depend on AFDCA effectiveness). We decided this score was problematic, because test scenarios were terminated at 30 minutes, before AF Strike could complete its mission. Therefore, point values observed for AF Strike were slightly negative, reflecting the fact that the simulation ended before AF Strike could accumulate its mission points. For this reason we considered only the AFDCA’s individual score and not any team-based score.

Figure 1 shows the raw total AFDCA score by the main conditions of the experiment. As one can see there is some indication of agent utility (higher scores with the agent available). This utility is noticeably more striking in the experienced group (n = 17) than in the inexperienced group (n = 20). However, the main effect of agent availability is marginally significant (F(1,35) = 5.01; p < .032), while the interaction is not (F(1,35) = 1.82; p < .186).

We also inspected some of our irrelevant or “nuisance” factors. These are factors that we would like to have little effect or be balanced over our experimental factors (agent availability and experience level). As these factors seemed adequately balanced across our experimental factors, we looked at the effect of these nuisance factors. These nuisance factors are time of scenario administration (i.e., practice effects for our two testing sessions) and geography used (Taiwan vs. Cyprus). There was no main effect of practice (i.e., no significant difference between scenario tested first or second; F(1,35) < 1; ns), indicating our training on the simulation interface was sufficient to bring participants to asymptote. However, there was a geography effect. Cyprus was “harder” in the sense that less points were earned in that scenario (130 vs. 143; F(1,35) = 30.13; p < .001).

The difference in scenario difficulty can also be seen by allowing an agent to play the AFDCA role (so there are no live players in the simulation). The agent’s AFDCA score in Cyprus would be 129 while in Taiwan it’s 152. Hence, if we wanted to compare an agent’s performance to the live players (i.e., average 129 and 152 to obtain a score of about 140), we would find that the agent performs better than the average WD in this study, but not too far from that average. Also we note that the agent’s performance is scenario dependent (i.e., WDs perform as well as the agent for Cyprus).

Given the scaling difference between point scores for Taiwan and Cyprus scenarios, we corrected the analysis implied by Figure 1, by dividing a subject’s score by 152, if that score derives from a Taiwan geography, or by 129, if that score derives from a Cyprus geography. What this does is express each subject’s score as a proportion of some standard performance, in this case the agent’s. With this ratio score which removed some uncontrolled variance, the effect of agent becomes more significant (F(1,35) = 5.80; p < .021), and the interaction suggested by Figure 1 also reaches significance (F(1,35) = 5.16; p < .029).



Recommendation Acceptance Rate

The simulation records recommendations it would have given under agent-available conditions regardless of whether the agent is allowed to issue such recommendations during the simulation. These recommendations and all human-issued orders are placed in a recommendations log file in chronological order. Hence, one can “define” whether a recommendation matches a human-issued order regardless of whether recommendations are actually available. Such matches occur if 1) a human order, issued later, matches a logged recommendation and 2) between the human order and the matching recommendation, no agent recommendation has superceded the initial matching recommendation. The second requirement says that the agent does not change its mind. This could occur if the agent subsequently recommends a different asset for the target or sends the asset from the initial recommendation after a different target. Unfortunately, counts of explicit accepts through the simulation interface (i.e., clicking on particular recommendations or clicking the “accept-all” button) was not available in these data, although future releases of the simulation will make this available. However, the current analysis provides an upper bound on the explicit acceptance rate.

Table 1 displays recommendation acceptance rates for TARGET orders as defined above. Specifically, TARGET orders “matching” TARGET recommendations are divided by the total number of TARGET orders for the main conditions of this experiment. On average about 1 in 6 TARGET orders matched an agent recommendation. There was little indication that explicitly presenting the recommendation (i.e., the agent-on condition) had any affect on match rates.

Table 1. Percentage of TARGET recommendations matching TARGET orders by experience level and agent availability.

	Agent Availability	
	ON	OFF
Experienced (n = 17)	15.6	18.9
Inexperienced (n = 21)*	15.2	13.5
*Note. Ns differ from benefit analysis owing to inclusion of partial data on one subject		

TANK orders were considered separately from TARGET orders, because TANK orders tend to be more ambiguous (e.g., recommending F15-A go to tanker does not necessarily invalidate an earlier recommendation to send F15-B to tanker). For TANK orders one can assess the impact of the agent by observing the frequency of “ran out of fuel” events for the main conditions of the experiment. Such events are logged in event log files for each WD. Table 2 presents this information in a coarse way.

Table 2. Frequency of WDs who had at least one “ran out of fuel” event logged for an asset under their care (by Experience Level and Agent Availability conditions).

	Agent Availability	
	ON	OFF
Experienced (n = 17)	0	6
Inexperienced (n = 21)*	7	7
*Note. Ns differ from benefit analysis owing to inclusion of partial data on one subject		

Experienced WDs appear to take advantage of refueling recommendations but the newly trained WDs do not. An interesting complexity in this data is that the raw number of TANK orders by condition does not replicate the trend in this table (i.e., Experienced are fairly flat across conditions; the Inexperienced group actually log more TANK orders when the agent is not available). A little reflection should clarify this. It is not the number of TANK orders but when they occur that is critical. Also nothing precludes a WD from issuing the same order twice (which will log the order twice), and this is not a rare occurrence. Some anxious WDs might have done so for TANK orders.

Agent and Weapon Director Performance: How Similar?

Table 1 would seem to suggest that the agent would have conducted itself quite differently than the live WDs did (i.e., about a 1 in 6 match rate on agent TARGET recommendations would seem to imply low similarity). However, there are other ways to assess similarity. Because the simulation allows an "autoplay" in which the live WD can be replaced by an agent playing his or her role, one can explicitly compare how the agent fought the war to some aggregated measure of how the WDs fought the war. "How to fight the war" might be defined as what resources were used to destroy what specific targets. In other words, one would catalog all agent asset-target-pairing choices and then compare that catalog to another such catalog that represents the WDs' pairing choices.

A scheme for representing the WDs' pairing choices in the "aggregate" would be to assess for each enemy (in a given scenario) what asset was most frequently chosen to intercept it. If the agent's choice matched the most popular WD choice, the agent was classified as matching the "aggregate" WD. In the case where there was more than one "most frequent choice" among the WDs' choices, the agent's choice was said to match if its choice was among the WD modes. Comparison in this fashion was done for each scenario and this data is summarized in Table 3.

Table 3. Frequency of "aggregate WD" to agent matches (maximum value is 60 matches) on the pairings of assets to enemy targets for the main conditions of the experiment.

	Agent Availability	
	ON	OFF
Experienced (n = 17)	45	30
Inexperienced (n = 21)*	44	41
*Note. Ns differ from benefit analysis owing to inclusion of partial data on one subject		

These data are interesting in that they suggest that the Experienced WDs are more similar to the agent (in terms of their choices) when the agent's recommendations are visible. Hence greater similarity to the agent would appear to be causing better performance (as experienced WDs have better performance when the agent's recommendations are visible). Unfortunately, mapping high vs. low agent similarity into high vs. low simulation performance is problematic. Other cells of Table 3 do not support this. We return to this issue, as well as a more complete description of this particular similarity score, in the disscusion.

Questionnaire Data: WD Impressions of the Agent.

Table 4 (below) gives a sense of WD subjective opinion for the value and usefulness of the agent. The scale goes from 1 to 6, with 1 meaning, "strongly disagree" and 6 meaning, "strongly agree." Ratings regarding the agent are not highly positive, but this may not be so surprising given the experimental and competitive nature of the technology evaluated. Hence, some notably lower ratings involved trust (e.g., item 49, 91) and similarity to self (i.e., the "expert WD" item 46). Other questions, not implying heavy reliance on the agent in life and death situations or that don't involve direct comparisons to the self show more ambivalent ratings (e.g., 53, 54, 56, 63).

However, other questions indicated that WDs could see the potential worth of the agent technology. For instance, WDs thought the agent could conceivably make them perform better (items 47, 67). Item 43 could mean that participants appreciated being able to launch a mission with a single mouseclick, provided the agent-suggested mission corresponded with their own. Finally, item 71 and 88 show that when the question is framed in terms of the agent as an additional tool, ratings for the agent are highest. Conversely, whenever the question can be interpreted as suggesting that the WD would have done better with the agent than without it, the agent rating generally decreases (e.g., item 77). In summary, WDs are willing to view the agent only as an efficiency booster (i.e., getting to the same level of performance easier) and not as a performance booster (i.e., getting a higher performance with than without the agent).

Table 4: Intelligent Agent Questionniare: Item Descriptives

Item#	Statements about the agent	Mean	Std. dev
43	The agent was easy to use	4.84	.95
44	The agent improved my performance	3.34	1.28
45	The display of the agent recommendation facilitated my performance	3.53	1.31
46	The agent behaved like an expert WD	2.47	1.16
47	The agent provided quality information	3.62	1.14
48	The agent's recommendations for actions were similar to my own	3.34	1.10
49	I trusted the agent's recommendations	2.82	1.25
50	I am solely responsible for my performance	4.58	1.22
51	Use of the agent was unnecessary	3.54	1.43
52	The agent made recommendations on high priority tasks (i.e., targets, RTB, TANK order)	3.68	1.44
53	I liked working with the agent	3.47	1.20
54	The agent decreased my workload	3.63	1.30
55	The display of the agent recommendation disrupted my task performance	3.32	1.25
56	The agent provided information necessary for me to complete my tasks.	3.45	.92
57	I was cautious in relying on the agent's recommendation	4.51	1.37
58	The rationales the agent provided for its recommendations were similar to my own	3.22	1.05
59	The agent influenced my decisions	3.26	1.22
60	I cannot be held responsible for actions I took based on agent recommendations	2.26	1.48
61	The agent decreased my control over my actions	2.13	1.14
62	The agent's recommendations narrowed my search for alternative actions	2.61	1.41
63	If I were running through training scenarios again, I would want to use the agent	3.32	1.45
64	The agent provided recommendations at opportune times	3.42	1.43
65	The agent's recommendations were predictable	4.11	1.25
66	The agent provided consistent information	4.13	1.07

67	I am skeptical of the agent’s capability to assist me	3.00	1.12
68	The agent’s information was consistent with my own	3.24	1.08
69	I accepted agent recommendations without reviewing the rationales	2.29	1.18
70	Responsibility for my actions is shared with the agent	1.89	1.03
71	The agent provided an additional resource in managing my tasks	4.42	1.29
72	The agent reduced my flexibility in performing my tasks	2.08	.78
73	I liked using this agent	3.61	1.42
74	On the job, the agent would be helpful in completing my tasks	3.63	1.20
75	I was willing to accept the agent’s recommendations during risky situations in the scenario	2.63	1.22
76	I had no choice but to follow the advice of the agent	1.39	.55
77	These tasks could have been completed as well without using the agent	4.24	1.13
78	The agent improved my flexibility in performing the tasks	3.5	1.18
79	The agent is useful to a novice WD	3.89	1.62
80	I was willing to accept agent recommendations during non-risky situations in the scenario	4.00	1.43
81	I am **not** solely responsible for my performance	2.24	1.50
82	The agent presented options I otherwise would not have considered	2.76	1.16
83	The agent is useful to an expert WD	3.08	1.38
84	The agent was dependable	3.32	1.07
85	I would trust the agent to perform certain tasks on my behalf	3.21	1.34
86	I am completely responsible for actions I took based on agent recommendation	4.47	1.59
87	The agent constrained my consideration of alternative actions	2.24	1.00
88	The agent’s recommendations were not useful in completing my tasks	2.42	.79
89	I was willing to accept the agent’s recommendations during uncertain situation in the scenario	2.45	1.18
90	The agent responded consistently to similar circumstances at different points in time	3.76	.94
91	I would trust the agent to perform the tasks of a WD	2.08	1.19

Discussion

Here we briefly summarize results and include any qualifications that may be necessary to interpret these data. In summary, many of our initial expectations were refuted. However, despite that, there was some evidence for the utility of having a decision aid – or agent recommendations – available. This appeared greatest in the Experienced group, counter to our intuitions. We also note that experience level did not affect performance in the task when the agent was not available. One might conclude from this that the level of competency in our participants was uniformly high and/or that the scenarios themselves were not sensitive to experience level (a ceiling effect).

On the other hand, even if the scenarios alone did not differentiate experience levels, experience did matter with regard to benefit from the agent. This suggests that the recommendation interface was more difficult for the relatively inexperienced WDs. Obviously ease of accepting a recommendation was not an issue. Instead the effort of evaluating all the presented recommendations was probably the critical difference between the experience groups. The Likert ratings showed the WDs to be conservative about accepting agent recommendations. In fact, WDs may have preferred to ignore recommendations, if they couldn't concurrently evaluate them while doing their primary job of evaluating the tactical situation themselves. The experienced group should be able to consider the agent more fully, because experts should be able to assess the tactical situation more quickly than the newly-trained WDs. Experts could have also been better at selectively accepting certain kinds of recommendations (e.g., refueling recommendations) for the same reason.

With regard to the overall benefit that was observed from exposure to the agent, we note that participants were not provided with any specific training on the logic and decision rules that drive agent recommendations, or with any particular agent feature. For example, the agent can easily "do the math" regarding fuel level and fuel consumption, and because of this, may recommend aircraft for longer distances than are comfortable for the WDs. Had the WDs been informed of this, the observed benefit of an online agent could have been higher. Parenthetically, we note that we were intentionally non-informative about the agent to minimize the potential for leading the subject.

Given an agent benefit, what primarily is the cause? This would seem a straightforward question, but for Table 1. That table shows the benefit cannot be explained by a greater tendency to accept recommendations when recommendations were available. However, Table 1 pertains only to TARGET orders and not to TANK orders. Table 2 suggests that the Experienced group may have received the benefit from the agent by avoiding "ran out of fuel" events (while the Inexperienced group did not).

One could also cite Table 3 as suggesting that greater similarity to the agent lead to better scores for the experienced WDs. However, similarity to the agent does not explain why experienced and inexperienced WDs do about the same when the agent is turned off (see Figure 1). If similarity to the agent had "caused" higher scores, we would have expected the Inexperienced group (41 matches to the agent) to outperform the Experienced group (30 matches to the agent) in the agent-off conditions. This did not (significantly) occur.

Table 3 is also complicated by another factor. All matches recorded do not reflect intercepts that were explicitly ordered by the WD (or the agent playing the WD's role). A large fraction of the intercepts reflect "targets of opportunity" taken by an ordered asset enroute to or after its primary intercept mission. It is not clear to what extent this compromises the agent-to-WD similarity measure implied by Table 3. This being the case, the explanation for agent-benefit implied by Table 2 can be preferred.

Other data relevant to agent/WD similarity are the WDs' ratings of the agent behavior relative to an "expert WD" (presumably themselves). Participants thought the agent did not behave like an expert. This rating may not mainly refer to the predictability of the pairings (i.e., see items 65 and 66, agent recommendations were predictable), but could refer to the timing at which such pairings occurred. Commits of assets to targets were recommended earlier than most of the WDs would have preferred, as this was mentioned in responses to the open-ended questions about agent behavior.

Other aspects of dissimilarity were gleaned from the process of building Table 3. One aspect is the tendency for WDs to refuel a low-fuel asset immediately at the beginning of a scenario, even though the asset (given the reach of its weapons) had enough fuel for a short mission. This difference (also mentioned in the open-ended responses) is interesting, because later in the scenario, when workload is high, WDs are ignoring refueling recommendations (Table 2). Finally, agent and WD behavior differed in the choice of F16s over F15s for some intercepts. The WDs avoided choosing F16s, which is consistent with AWACS doctrine, because F16s are used less in air-to-air missions. This bias against F16s (which future versions of the agent could be brought in line with) was not detectable in the open-ended responses.

Conclusions and Future Issues

We found a complicated series of results that indicated some benefit for the agent, in spite of relatively conservative agent usage. The fact that the benefit is greater for the WD experts implies that only the experts can manage the consideration of many simultaneously presented recommendations. Simultaneous presentation of as much as six or more recommendations could occur, depending on the work load at a given point in the scenario.

The number of recommendations presented during a simulation tick (i.e., 10-second radar sweep) could relate directly to a WD's objective (and perceived) workload at that time. If the agent becomes the most "vocal" during times of high-workload, conservative usage may not be so surprising. WDs ascribe principle responsibility for tactical assessment to themselves and may want to ignore advice when they are very busy. It is a challenge to the interface to make recommendation-evaluation less of a "dual-task" (along with assessment) for the inexperienced WD. Obvious possibilities for improving the interface would involve prioritizing recommendations so that only the most time-critical ones (up to a certain number, say 3) were presented during a tick. In this regard refueling recommendations would routinely be presented above others (other things equal), as they are time-critical by definition. Another possibility, which could be added on top of prioritization, would be to more strongly embed the recommendation in the context of normal tactical assessment performed by the WD. As an example, whenever a WD brought up an information window on an asset (to inspect fuel and armament), a recommendation for that asset could be displayed (if one existed) along with the requested information. Otherwise the recommendation would not be displayed. Finally, the possibility of tailoring the "vocalness" of the agent so that operator could choose how many recommendations to see at any given time should also be investigated.

An important limitation of any specific study of intelligent agents is that there are many different ways of testing "agent" technology, and we only chose to focus on one of them, namely as a decision aid. The relationship between number of agent recommendations and workload suggest an independent way of validating the agent (and also suggests an interesting use of agent-technology in standard simulation conditions--i.e., as a valid measure of workload when there are no decision aids available). Agent technology is also directly validated by measures of the perceived realism of the synthetic forces. In fact, the technology behind the agent decision aid is the very same technology governing the "intelligence" of the computed-generated forces in this specific case. Therefore agents can also be evaluated by how realistically an agent-directed teammate or enemy force acts (either by a subjective rating or a "Turing" test).

Finally, much more research is needed to assess the issue of training benefit from simulation experience with agent-managed forces (both enemy and friendly) as compared to similar experience with live players. The possibility of less manning and coordination per training opportunity is a major attraction for this technology. While training the cognitive aspects of the job domain should be comparable (or even better) in the context of agent-managed forces, other source of variance, not yet in the domain of agents, may suffer. For instance, one of the more interesting (and ambitious) challenges to this technology is whether effective inter-team management skills can be fostered with agent-driven teammates.

References

Coovert, M., Riddle, D., Gordon, T., Miles D., Hoffman, K., King, T., Elliott, L. R., Schiflett, S. G., & Chaiken, S. (2001). The impact of an intelligent agent on weapon directors behavior: issues of experience and performance. Proceedings of the 2001 Command and Control Research and Technology Symposium, Annapolis, MD.

Coovert, M. D., Gordon, T., Foster, L. L., Riddle, D., Miles, D. E., Hoffman, K. A., & King, T. S. (1999). AWACS weapons directors cognitive/behavioral categories: The final report of USF's AWACS research group. Technical Report. University of South Florida.

Coovert, M. D., & Riddle, D. (1999). The application of rough set theory to AWACS weapons directors C3STARS data: Final report. Technical Report. University of South Florida.

Dalrymple, M. A. (1991). Evaluating Airborne Warning and Control System strategy and tactics as they relate to simulated mission events. (AL-TP-1991-0049, AD A242820). Brooks AFB, TX: Armstrong Laboratory.

Elliott, L. R., Schiflett, S. G., Hollenbeck, J. R., & Dalrymple, M. (In press). Situational awareness and team performance in realistic command and control scenarios. In M. McNeese, E. Salas, & M. Endsley (Eds.) Group Situational Awareness: New Views of Complex Systems. Human Factors and Ergonomic Society Press.

Elliott, L. R., Chaiken, S., Dalrymple, M., Petrov, P., & Stoyen (2000). A. Simulation-based agent support in a synthetic team-based C2 task environment. Proceedings of the 2000 Command and Control Research and Technology Symposium, Monterey CA.

Fahey, R. P., Rowe, A., Dunlap, K. & DeBoom, D. (2000). Synthetic task design (1): preliminary cognitive task analysis of AWACS weapons director deams. (AFRL-HE-AZ-TR-2000-0159). Mesa AZ: Air Force Research Laboratory, War Fighter Training Research Division.

Gordon, T., Coovert, M., Riddle, D., Miles D., Hoffman, K., King, T., Elliott, L. R., Schiflett, & Chaiken, S. (2001). Classifying C2 decision making jobs using cognitive task analyses and verbal protocol analysis. Proceedings of the 2001 Command and Control Research and Technology Symposium, Annapolis, MD.

Klinger, D. W., Andriole, S. J., Militello, L. G., Adelman, L., Klein, G., & Gomes, M. E. (1993). Designing for performance: A cognitive systems engineering approach to modifying an AWACS human-computer interface. (AL/CF-TR-1993-0093, AD A275187). Wright-Patterson AFB, OH: Armstrong Laboratory.

Petrov P., Stoyen A., & Myers G. (2000)* AWACS Weapons Director Trainer Tool: USAF SBIR Phase II Final Report. Contract: F41624-98-C6011, 21st Century Systems, Inc.

Schiflett, S. G. & Elliott, L. R. (2000). Synthetic team training environments: Application to command and control aircrews. In D. Andrews and H.R. O'Neil, Jr. (Eds.), Aircrew training and assessment (pp. 311-335). Mahwah NJ: Lawrence Erlbaum Associates.

* The contract report is proprietary and not available to the general public, but information on more evolved versions of the WD-IAA task, described herein, may be available at www.21csi.com by inquiring about the AWACS-AEDGE (tm) product.