Visualizations for Communicating Intelligent Agent Generated Courses of Action

Jessica Bartik[1(✉)], Heath Ruff[2], Gloria Calhoun[1], Kyle Behymer[2], Tyler Goodman[1], and Elizabeth Frost[2]

[1] Air Force Research Laboratory, 711 HPW/RHCI, Dayton, OH, USA
{Jessica.Bartik.1,Gloria.Calhoun,
Tyler.Goodman.3}@us.af.mil
[2] Infoscitex, Dayton, OH, USA
{Heath.Ruff.ctr,Kyle.Behymer.1.ctr,
Elizabeth.Frost.6.ctr}@us.af.mil

Abstract. Future human-autonomy teams will benefit from intelligent agents that can quickly deliberate across multiple parameters to generate candidate courses of action (COAs). This experiment evaluated the design of an interface to communicate agent-generated COAs to a human operator. Twelve participants completed 14 trials, each consisting of a series of tasks that required participants' selection of the best COA in terms of quality, speed, fuel, and detectability parameters. Trial score and speed of participants' selection were measured as a function of COA visualization (1, 4, or 8 COAs) as well as the type of agent. Supplemental trials in which participants could choose which visualization to employ for COA selection were also conducted. The data showed that presenting multiple COAs were better than a single COA. Differences between the 4 and 8 COA visualizations were not quite as definitive: selections were significantly faster with 4 COAs than 8, but participants' preferences were divided based upon agent comprehensiveness and individual strategy differences. The results also showed that the agent's reasoning process should be communicated more precisely besides just what parameters are being considered in generating COAs.

Keywords: Human-autonomy teaming · Intelligent agent · Parallel coordinates plot · Plan comparison · Visualization · Human autonomy interaction

1 Introduction

Given the increasing pervasiveness of adversarial threats and the stable Department of Defense (DoD) workforce numbers, autonomous technologies are being developed that enable a single operator to control multiple autonomous vehicles. One example of such technologies is an intelligent agent that uses a cognitive domain ontology to categorize situations and develop multiple courses of action (COAs) in response to mission events [1]. The agent is capable of analyzing and ranking all of the potential COAs according to multiple optimization criteria for a given high level goal/task. For example, the agent

This is a U.S. government work and not under copyright protection in the U.S.;
foreign copyright protection may apply 2019
J. Y. C. Chen and G. Fragomeni (Eds.): HCII 2019, LNCS 11575, pp. 19–33, 2019.
https://doi.org/10.1007/978-3-030-21565-1_2

could determine the vehicle most likely to find a target and expend the least amount of time and fuel in the process.

A significant challenge facing interface designers is determining how to best present agent-generated COAs to an operator, as each one imposes information retrieval costs. Moreover, presented alternatives can potentially mask aspects of the problem space and influence the operator's decision-making [2]. One approach is to present a single solution to the operator, the solution that the agent has determined to be the best. Another approach involves modeling to generate alternatives [3], focusing on generating a small set of alternatives that are "good" in terms of achieving the operator's goal but different in respect to the relevant parameters of the solution space. This approach aims to generate options that can achieve the commander's objective but vary in other parameters. For example, three COAs (A, B, C) might be generated based on the high-level goal of getting a vehicle to a specified location in under 20 min, but COA A minimizes fuel use, COA B maximizes stealth, and COA C minimizes future maintenance costs.

To address this challenge, an experimental testbed was developed in which participants were instructed to achieve the highest score possible (while also avoiding fuel violations and detections) in a specific time window by completing a series of COA selection tasks. For each task, eight possible COAs were generated. Each COA had four associated parameters—quality points, time, fuel, and detection. For example, selecting COA A might give the operator 25 quality points, take 16 min, cost 13 fuel units, and have a 50% chance of being detected, while using COA B might give the operator 10 quality points, take 5 min, cost 12 fuel units, and have a 17% chance of being detected. To enable participants to compare the eight COAs across the four parameters, the testbed included a parallel coordinates plot (see Fig. 1) referred to as the 'Vehicle Comparison Tool' [4]. This tool made the tradeoffs between COAs immediately visible. For example, in Fig. 1, COA G (colored orange) was the highest quality, but consumed the most fuel. COA D (colored green) was the second lowest quality, but was least likely to be detected by enemy forces.

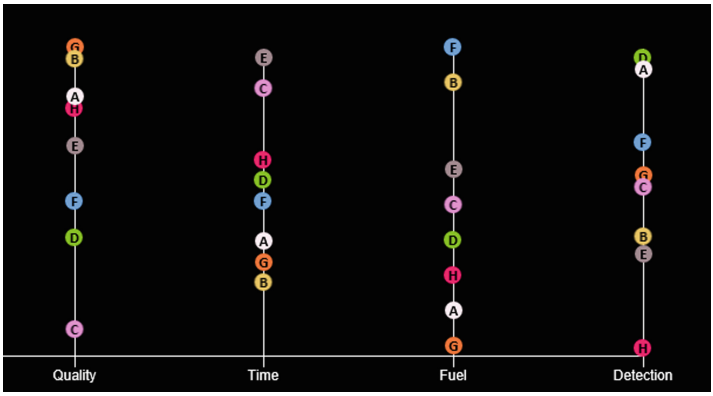

Fig. 1. Vehicle Comparison Tool representing eight COAs across four parameters. For each parameter, higher was better (e.g., COA C would generate the least amount of quality points, COA E would take the least amount of time, COA F would use the least fuel, and COA D had the smallest chance of being detected). (Color figure online)

Using this testbed, an experiment was conducted [5] in which participants were tasked with evaluating and selecting the "best" of eight simulated agent-generated COAs. Four visualizations were evaluated, varying in COA number and type: (1) a single COA (highest quality), (2) the four highest quality COAs, (3) the four COAs with the highest value for each parameter (the most quality points, the least time, the least fuel cost, and the least chance of detection), and (4) all eight COAs. Regardless of visualization condition, participants had the ability to call up parameter details of all eight COAs. Data from the experiment indicated that the single COA visualization was significantly less effective than the other visualizations. However, the results did not indicate a clear best option from the other three visualizations.

These findings may reflect limitations of the testbed. Although the testbed provided an engaging test environment that could be rapidly trained, the COA selection tasks did not have high attentional demands compared to operational stations. For instance, participants had as much time as they desired to drill down for COA details not presented, perhaps mitigating differences between visualizations. An additional limitation may have been the lack of a "best COA" for each task. Instead, the eight COAs were balanced across the four parameters making it ambiguous as to which COA was the best option for a specific task. This technique may have led to the lack of significant differences in objective performance measures. Including an agent that could reason across different parameters for the given experimental task and provide participants with one or more recommendations could combat this effect. For example, in one visualization an agent could reason across the four parameters and recommend a single "best" COA. In another visualization, the agent could return the top four solutions.

As such, a follow-on experiment was conducted that employed two intelligent agents. Each was capable of reasoning across multiple parameters to generate and rank COAs, but differed in the number of parameters considered. The testbed and training were also slightly modified to impose more temporal demands. Participants were briefed that their COA selection time was limited to 30 s. To help manage the limited selection time, a digital readout that counted down was added to the testbed. Besides creating a more temporally demanding test environment, these modifications added an interface element that also increased requirements to shift attention. The experiment reported herein utilized this enhanced version of the testbed to examine which of three visualizations was most effective in aiding participant COA selection performance, as well as the impact of agent reasoning comprehensiveness.

2 Method

2.1 Participants

Twelve volunteer employees working at a U.S. Air Force Base between the ages of 19–51 (M = 35.08, SD = 10.68) participated in the study. All participants reported normal or corrected normal vision and color vision.

2.2 Experimental Design

Six conditions were evaluated, varying in agent type and visualization (see Table 1). The more comprehensive Agent A reasoned across all four parameters (quality, time, fuel, detection) when generating and ranking COAs for a given task. Agent B only reasoned across three of the four parameters (quality, time, and fuel), excluding detection from its reasoning.

Table 1. Illustration of the six experimental conditions

	Agent A	Agent B
Visualization	1 COA	1 COA
Visualization	4 COAs	4 COAs
Visualization	8 COAs	8 COAs

Trials were blocked by agent type and visualization. For example, a participant completed three blocks of trials (i.e., one per visualization) with one agent and then three more blocks of trials with the other agent. For each participant, the order of the visualization blocks with each agent was the same. However, across participants, the order of the visualizations and two agent types was counterbalanced. Within each of the six blocks, participants completed two trials. For each 480 min (simulated) trial participants were presented a series of generic tasks (i.e., tasks lacked context and were simply labeled "Task 1", "Task 2", etc.) with the number of tasks per trial being dependent upon participants' COA selections (ranged from 12 to 21, $M = 14.65$, $SD = 1.55$). To complete each task, participants were trained to select one of the eight COAs (i.e., vehicles) labeled A–H within 30 s. Each COA differed on four parameters: quality (i.e., the number of points the participant could accumulate for selecting this COA), time to complete the task using this COA, fuel used, and probability of detection.

In addition to the series of experimental trials described above, participants completed two supplemental trials, one with each agent type (order counterbalanced across participants). For each of the tasks within these trials, participants selected which visualization (1, 4, or 8 COAs) they wanted to have presented (and they could switch as much as they desired between the different visualizations). Unlike the prior experimental trials, participants had an unlimited amount of time to make decisions.

2.3 Test Stimuli

Figure 2 illustrates how the same data set was depicted in the Vehicle Comparison Tool for each of the six conditions. Panes 1–3 show COAs suggested by Agent A. In (1), the top COA (F) is shown. In (2), the top four COAs are shown (adding B, D, and E to F). In (3), all eight COAs are shown. COAs suggested by Agent B are shown in Panes 4–6. In (4), the top COA (E) is shown. In (5), the top four COAs are shown (adding B, F, and G to E). In (6), all eight COAs are shown.

A range of possible values was determined for each of the four parameters and assigned to the eight COAs according to the method outlined in [5]: quality 10–60 points, time 15–45 min, fuel 5–22 gallons, and detection probability 5%–60%. Upon COA selection, a random number between 1–100 was generated. If that number was below the chosen COA's detection probability, the selection would result in a detection. If the number was above the chosen COA's detection probability, the COA selection would not result in a detection and the quality points for that task would be awarded. For example, if a COA with a 20% chance of being detected was selected and the random number generated was "19", the result would be a detection.

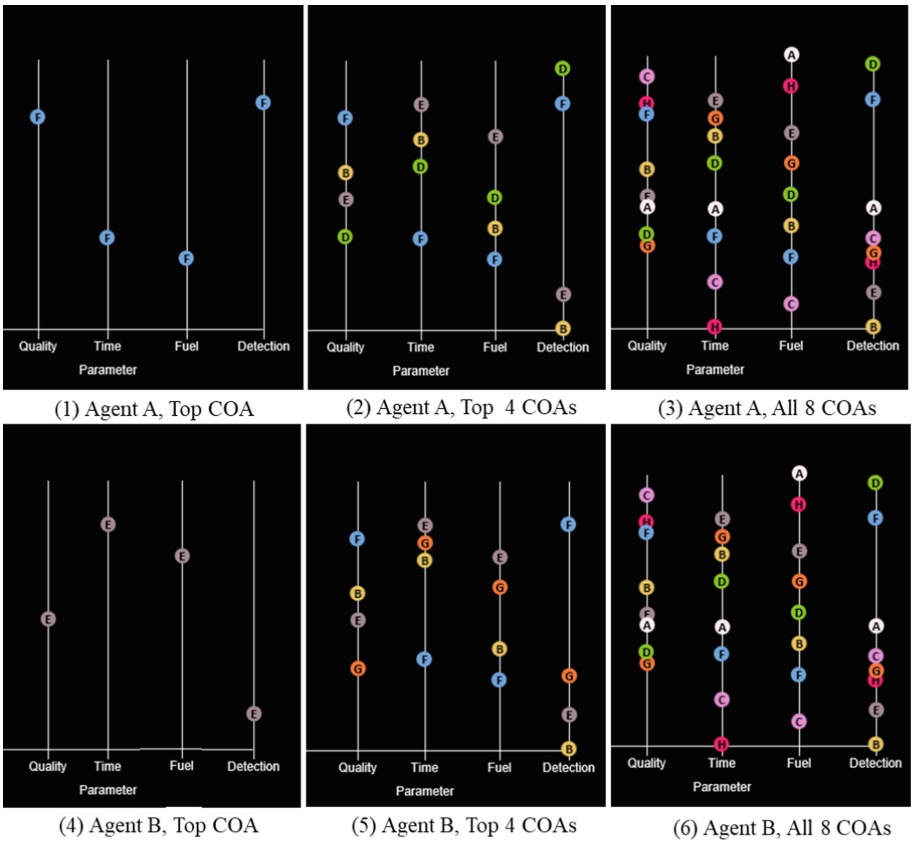

Fig. 2. Six experimental conditions differing in agent type (A or B) and visualization (1, 4, or 8 COAs)

Regardless of agent type or the number of COAs shown, participants were provided the top agent COA for each task via a text readout located directly below the Vehicle Comparison Tool (e.g., "The agent recommends Vehicle C"; see Fig. 3). Participants also were provided the ability to 'drill down' to see the associated values for all eight COAs. Two drill down methods were available (see Fig. 3 for the results of both

methods). One allowed participants to use a mouse to hover over a parameter in the Vehicle Comparison Tool to see the values of each of the eight COAs for that parameter. The second method allowed participants to hover over a COA/vehicle button to see that vehicle's values for each of the four parameters.

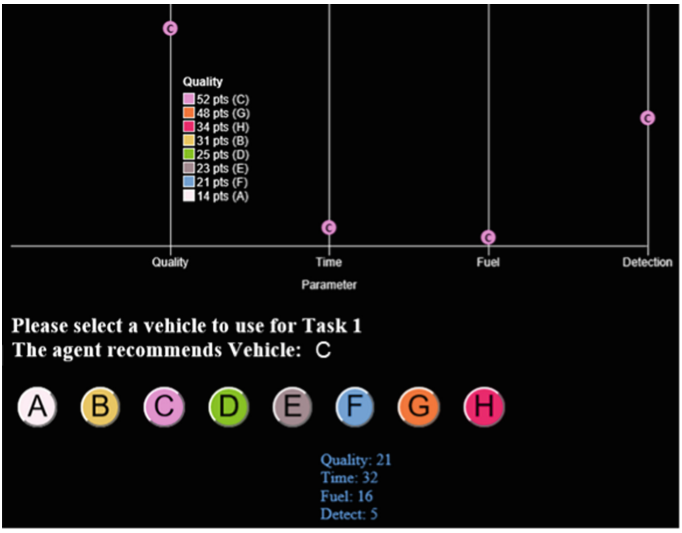

Fig. 3. Close up view of a portion of the testbed showing two drill down methods and agent recommended COA text readout. Hovering over a parameter (e.g., Quality) showed the parameter's specific value for all eight COAs, from highest to lowest. Hovering over a vehicle selection button (e.g., F) showed the values of all four parameters for that specific COA.

2.4 Trial Procedure

Participants were seated in front of the testbed (see Fig. 4) that was presented on a 24 inch monitor. A mouse was used for inputs. To start each experimental trial, participants clicked a button on the monitor labeled *BEGIN*. Participants were trained to use the *Vehicle Comparison Tool* (see Fig. 4) to determine which COA to select for each task within the trial. Participants indicated their selected COA/vehicle by clicking on the corresponding lettered circle (see *Vehicle Selection Buttons* in Fig. 4). The time limit for making each selection was 30 s. A digital readout that started at 30 s for each task counted down until either the participant made a selection or time had expired.

Upon COA selection, the *Scoreboard* (see Fig. 4) updated based on the results of the participant's selection. *Score* increased by the chosen COA's associated quality value if the vehicle wasn't detected. If detected, *Score* remained the same and *Detections* increased by one. *Time Remaining* decreased based on the amount of time associated with that COA selection. If participants took more than 30 s to make a decision *Time Remaining* decreased by an additional 25 min. Similarly, if participants had to refuel a vehicle, *Time Remaining* decreased by 25 min. The *Fuel Status Display*

(see Fig. 4) updated to reflect the amount of fuel used. If selecting a COA resulted in a vehicle having less than 50 gallons of fuel, *Fuel Violations* increased by one. Participants continued to receive new tasks until there were no COAs whose associated time requirement was less than the *Time Remaining*, at which point the trial ended.

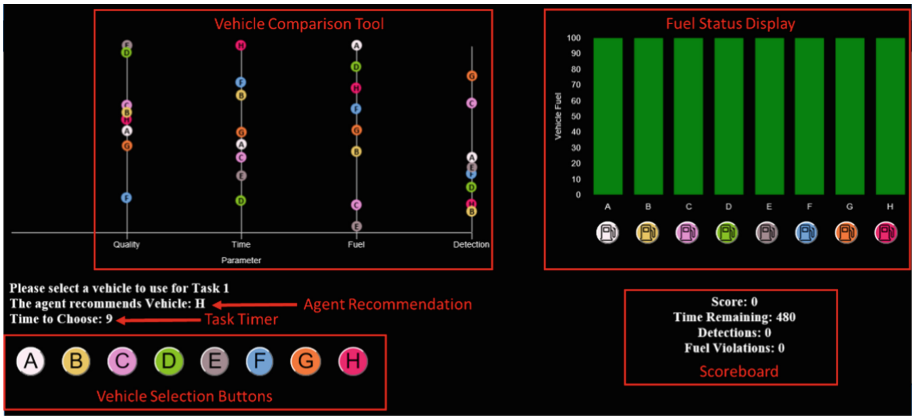

Fig. 4. Screenshot of the experimental testbed with key elements annotated

The supplemental experimental trials followed a similar procedure but there was no timer and participants could use the buttons (labeled 1, 4, and 8 in Fig. 5) to choose, as well as flip between, the visualization(s) presented for each task.

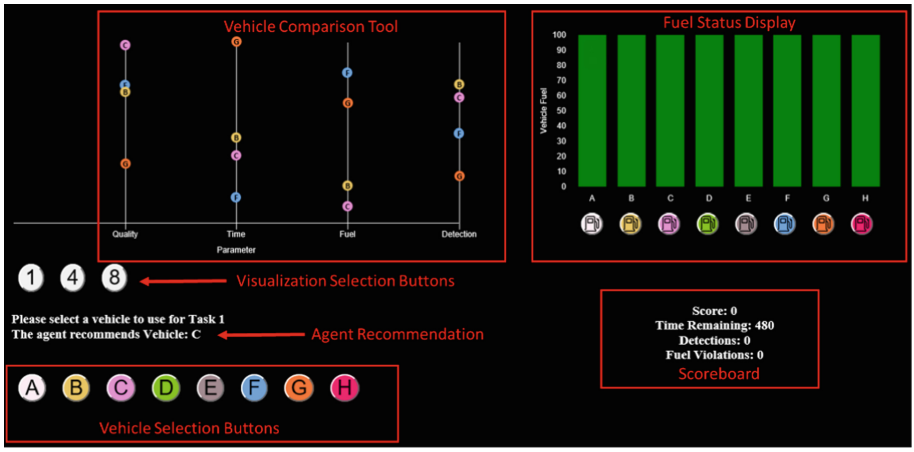

Fig. 5. Screenshot of the experimental testbed configured for the supplemental trials with key elements annotated

2.5 Test Sessions

Upon arrival, participants read the informed consent document and completed a demographics questionnaire. Next, participants were trained on the COA selection task, beginning with a discussion of the four parameters they needed to consider when selecting a COA. Participants learned they had four equally important goals: (1) maximize score, (2) avoid detections by enemy forces, (3) avoid fuel violations, and (4) avoid timeouts. Participants were then briefed on the major testbed components for the experimental trials to include the Vehicle Comparison Tool, Agent Recommendation, Timer, Fuel Status Display, Vehicle Selection Buttons, and Scoreboard (Fig. 4). Finally, participants were trained on the four visualizations, two agent types, and both drill down methods. After conducting a series of training trials and being trained to a level deemed appropriate by the experimenter, the participants began the experimental trials. A similar training procedure was used for the supplemental trials, highlighting the changes in the testbed and requirement to select a visualization at the start of each task.

After each block (6 total), participants were given a Post-Block questionnaire asking about their perceived performance, workload, ability to identify the best COA, ability to identify the best COA for a specific parameter, how often they drilled down, and the strategy they used for the visualization they just experienced. All questions used a five point Likert scale, with the exception of the strategy question, which was open-ended. Following the completion of three blocks of trials with each agent, a Post-Agent questionnaire was administered asking the participants to rank the visualizations from 1–3 (1 = best, 3 = worst) in terms of performance, response time, accuracy, and trust. After trials with all six conditions were finished, participants completed a Post-Experiment questionnaire containing open-ended questions asking participants which visualization they most and least preferred, whether or not their strategy changed depending on the visualization, and which drill down method they preferred. There were also prompts for suggestions to enhance the symbology and elements of the testbed.

Lastly, after the supplemental trials, a questionnaire was administered with open-ended questions asking which visualization participants tended to select most often to complete the COA selection tasks and whether that tendency changed based on agent type. They were also queried on whether they preferred having the ability to call up a different number of COAs versus having a default number of COAs in conjunction with drill down capabilities. Total session time, per participant, was approximately 2 h.

3 Results/Discussion

Data for each participant were collapsed across the two trials for each visualization. Objective data were analyzed with a repeated measures Analysis of Variance (ANOVA) model. An ANOVA model was also applied to the Post-Block questionnaire responses. The Post-Agent ranking data were analyzed using the Friedman nonparametric test of significance. Post-hoc Bonferroni-adjusted t-tests were performed for significant ANOVA and Friedman results. Error bars in figures are standard errors of the mean.

3.1 Performance

Score, response time (amount of time between visualization presentation and COA selection), and detection data were analyzed as objective measures of performance. The results of an ANOVA indicated that there was a significant interaction between agent type and visualization on mean score ($F(2,22) = 7.45$, $p = .003$, $\eta_p^2 = .40$). When utilizing the less comprehensive Agent B, participants scored lower with the 1 COA visualization, compared to when 4 COAs were presented ($t(11) = 3.80$, $p = .05$, $d = 1.35$, see Fig. 6). Rankings from the Post-Agent questionnaires were aligned with these results. Participants ranked their COA selections less accurate with the 1 COA visualization compared to the 4 COA visualization (Agent B) and the 4 COA and 8 COA visualizations (Agent A). Overall performance was also ranked worse with the 1 COA visualization versus the 4 COA visualization (Agent B) (see Tables 2 and 3).

These performance results are aligned with participants' ratings on their ability to make COA selections with the three visualizations. For example, their ratings indicated that it was more difficult to select the best vehicle when only 1 COA was presented compared to the 8 COA visualization ($t(11) = 4.45$, $p = .003$, $d = 0.44$). In contrast, differences in vehicle selection ability/performance between the 4 COA and 8 COA visualizations were not clear. In fact, participants' ratings were divided with respect to these visualizations. Seven of twelve participants indicated a preference for 4 COAs commenting that it provided a variety to choose from but was less cluttered and eliminated undesirable options, compared to 8 COAs. Another four participants preferred the 8 COA option, noting that presentation of more options provided a "big picture" and an ability to judge COAs just by their relative positioning on the axes of the Vehicle Comparison Tool, with less drilling down for exact numeric parameter values. Finally, one participant's preference depended on the agent in effect, desiring 4 COAs for the more comprehensive Agent A and 8 COAs for Agent B that didn't consider detection.

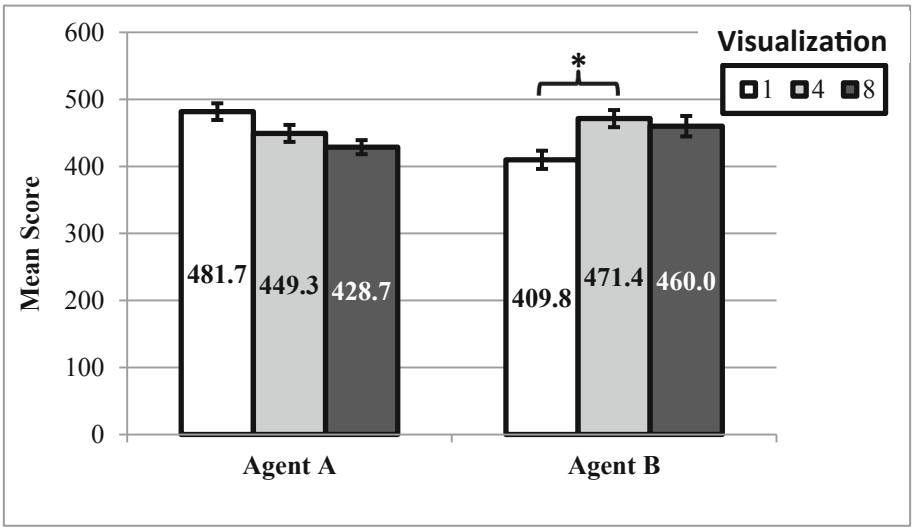

Fig. 6. Mean score by agent type and visualization

Table 2. Post-Agent questionnaire Friedman results

Variable name	$\chi^2(2)$	p	W
Performance (Agent B)	9.5	.009	.34
Accuracy (Agent B)	6.0	.05	.25
Accuracy (Agent A)	10.2	.006	.42

Table 3. Post-Agent questionnaire post-hoc results

	One COA		Four COAs		All eight COAs				
Variable name	*M*	*SE*	*M*	*SE*	*M*	*SE*	*t*(11)	*p*	*d*
Performance (Agent B)	2.67	0.19	1.42	0.15			5.00	.001	2.13
Accuracy (Agent B)	2.50	0.23	1.50	0.15			3.32	.021	1.48
Accuracy (Agent A)	2.75	0.13	1.58	0.19			4.84	.002	2.04
Accuracy (Agent A)	2.75	0.13			1.67	0.23	3.46	.016	1.70

Similarly, the mean number of detections differed as a function of agent type and visualization ($F(2,22) = 6.94$, $p = .005$, $\eta_p^2 = .39$). For the more comprehensive Agent A, participants' COA selection resulted in a higher number of detections when 8 COAs were presented compared to the other two visualizations: 1 COA ($t(11) = 5.61$, $p = .002$, $d = 1.69$) and 4 COAs ($t(11) = 4.53$, $p = .013$, $d = 1.13$, see Fig. 7). This finding was unexpected as it was anticipated that there would be more detections with Agent B, the agent that doesn't consider detections in its reasoning, than with Agent A. However, the fact that participants were trained on Agent B's shortcoming might have made them more acutely aware of detections for trials supported by Agent B, thus improving their performance. This training on the respective comprehensiveness of Agents A and B, coupled with comments suggesting that the 8 COA visualization presented "clutter" issues could explain why more detections were observed when 8 COAs were presented as compared to when 1 or 4 COAs were presented with Agent A.

With respect to participants' response time, agent type did not have an effect (p = .221). In contrast, a main effect of visualization on response time was found ($F(2,22) = 67.15$, $p < .001$, $\eta_p^2 = .86$, see Fig. 8). Participants responded significantly quicker with the 4 COA visualization as compared to both the 1 COA ($t(11) = 14.34$, $p < .001$, $d = 6.42$) and 8 COA visualizations ($t(11) = 4.38$, $p = .003$, $d = 2.10$). Additionally, mean response time was faster with the 8 COA visualization than the 1 COA visualization ($t(11) = 7.63$, $p < .001$, $d = 2.08$).

Subjective data supported these findings. For the visualization that presented only 1 COA, ratings and comments were generally unfavorable, citing that it failed to provide adequate information to base COA selection. This made it necessary to exercise the hover functionality repeatedly to call up information before selecting a COA, thus increasing response time. This is also probably the basis of the participants' workload ratings (main effect of visualization: $F(2,22) = 6.67$, $p = .006$, $\eta_p^2 = .38$) being higher with COA 1 than COA 4 and 8 ($t(11) = 2.93$, $p = .041$, $d = 1.06$ and $t(11) = 3.19$,

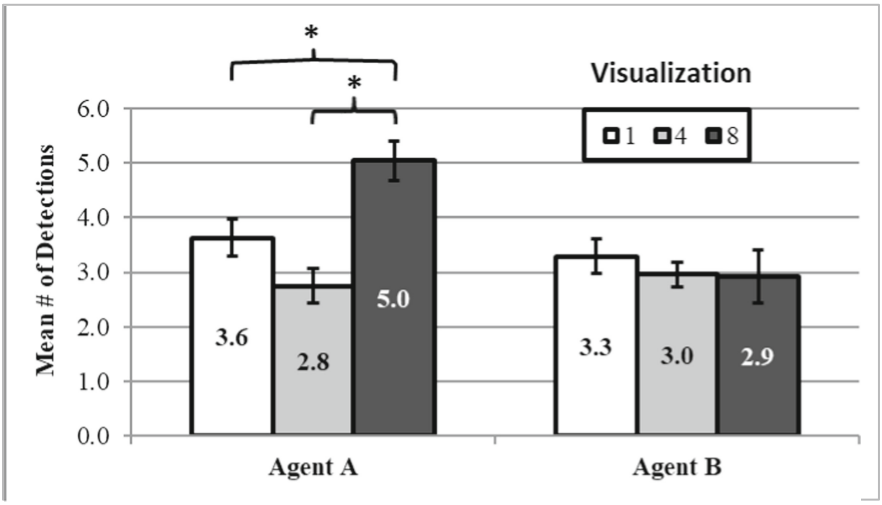

Fig. 7. Mean number of detections for each visualization by agent type

p = .026, d = 1.21, respectively). In contrast, the visualizations with 4 and 8 COAs provided more information.

Likewise, participants' frequency in using hovers to drill down for additional information significantly differed across visualizations ($F(2,22) = 9.52$, p = .003, η_p^2 = .46). The drill down functionality (with Agent A) was utilized more with 1 COA than with visualizations that had 4 COAs and 8 COAs ($t(11) = 3.08$, p = .031, d = 1.46 and $t(11) = 4.21$, p = .004, d = 2.23, respectively). Similar results pertaining to drill down frequency were shown in the ranking of the visualizations on the Post-Agent questionnaire ($\chi^2(2) = 12.17$, p = .002, W = .51).

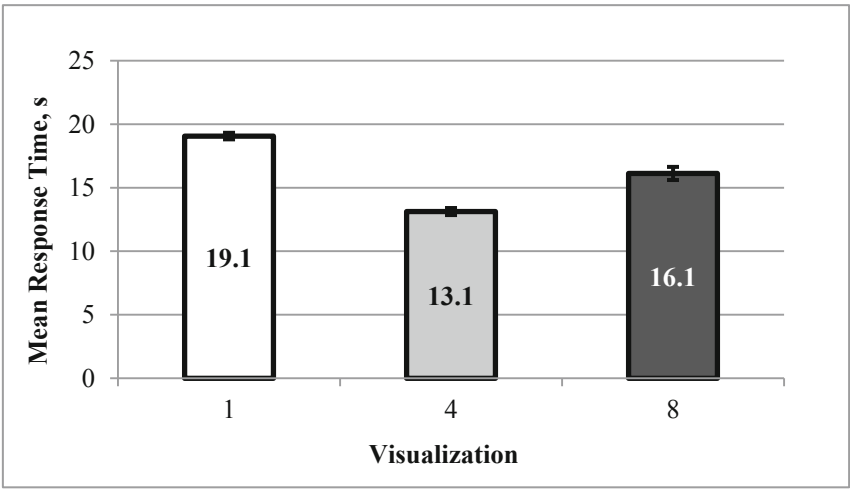

Fig. 8. Mean response time for each of the three visualizations

3.2 Agent Reliance

The term "Agent Reliance" is used here when reporting the ANOVA results pertaining to the mean percentage of tasks that participants' COA selection matched the agent's recommendation. To be clear, any matching response may just have been coincidental rather than reflect a tendency for the participant to depend or rely on the agent's recommendation. The results of the ANOVA showed that the mean percentage significantly differed between the two types of agents, as a function of visualization. The participants' selection and agent's recommendation were the same more frequently in the 1 COA visualization condition with Agent A compared to when the less comprehensive Agent B was used ($F(2,22) = 15.16$, $p = .001$, $\eta_p^2 = .58$; see Fig. 9). This was the case for all three visualization conditions with Agent B (1 COA, 4 COAs, and 8 COAs; ($t(11) = 5.27$, $p = .004$, $d = 1.81$; $t(11) = 4.68$, $p = .01$, $d = 1.08$; and $t(11) = 7.81$, $p < .001$, $d = 1.12$ respectively). As previously reported, the 1 COA visualization resulted in longer response times. As such, this effect could be explained by the participants' propensity to select the more comprehensive Agent A's recommended COA when the timer approached expiration in order to avoid the 25 min simulated time penalty associated with a timeout. Overall, participants completed COA selections before the timer expired on 85.42% of trials.

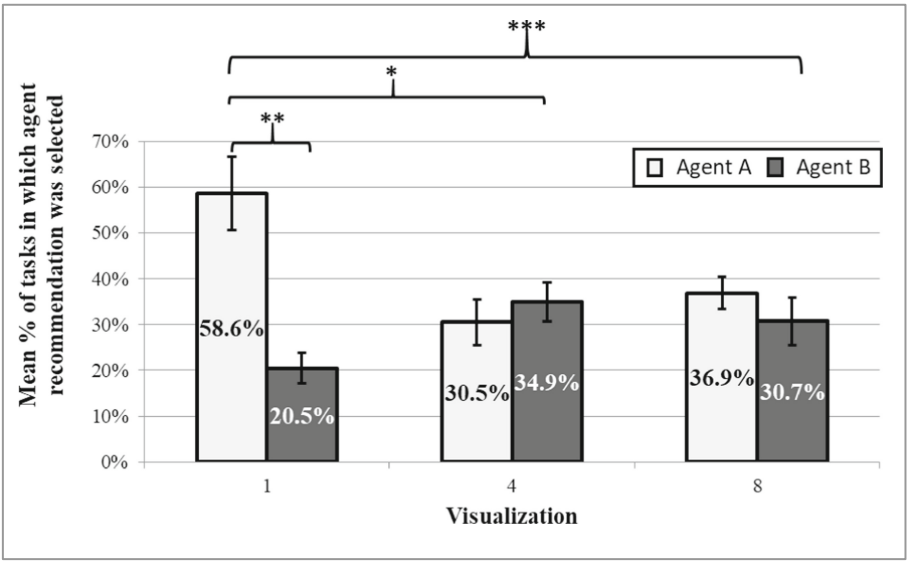

Fig. 9. Mean percentage of COA selections that matched agent's recommendation for each visualization

Despite this interaction, the overall low mean rate of reliance ($M = 35.35\%$) and frequency that participants used the drill down functionality suggest that participants did not trust either agent. Moreover, the subjective ratings regarding trust in the agents did not significantly differ. Instead, participants' comments indicated that many did not

rely on either agents' recommendations (e.g., "I didn't trust either agent to rely on them"). Even for Agent A that considered all four parameters, participants raised questions on the degree to which the agent's algorithm took each parameter into account (e.g., several commented that the agent weighted fuel remaining too high). The recorded comments imply that participants' knowledge of which parameters were considered by the agent was inadequate – many wanted more transparency into the agent's processing (e.g., the relative weights and parameter thresholds each agent took into account in coming up with the recommended COA).

3.3 Visualization Usage in Supplemental Trials

The supplemental trials provided data on participants' preference when they were free to choose and switch between the 1, 4 and 8 COA visualizations. In the questionnaire data, nine of the twelve participants indicated they preferred having the option to change the number of COAs presented. The results of an ANOVA examining actual selections made by the participants revealed a significant effect of agent type on visualization choice ($F(1,11) = 6.61$, $p = .03$, $\eta_p^2 = .40$, see Fig. 10). (This ANOVA did not include instances when the single COA option was selected ($n = 1$)). The 4 COA visualization was chosen more often by participants for trials with the more comprehensive Agent A. With Agent B, the 8 COA visualization was chosen more frequently. It is possible that participants felt they needed more information with the less comprehensive Agent B to make a decision and thus they preferred the 8 COA visualization's more detailed information over having only 4 options.

The number of times participants changed the visualization for each task was also examined. The difference between the number of changes with Agent A and Agent B was not significant ($t = .11$) and ranged from 0 to 33 times per trial ($M = 7.50$, $SD = 8.80$).

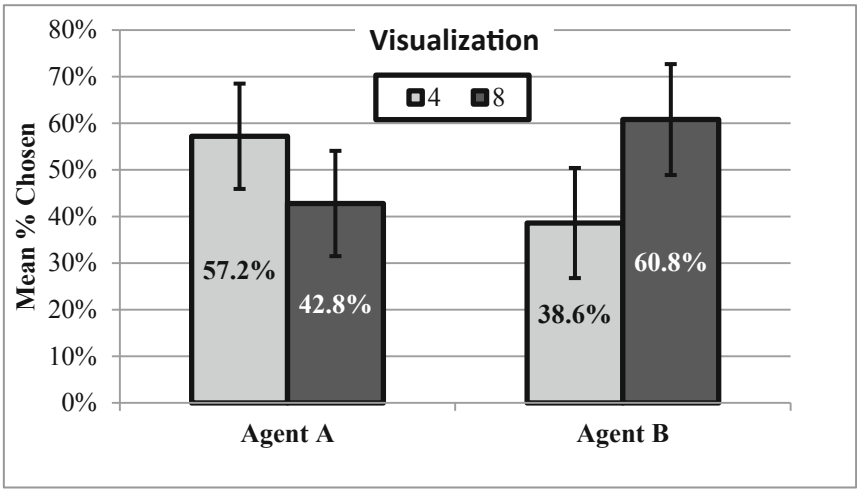

Fig. 10. Mean percentage of tasks in which participants chose the 4 COA and 8 COA visualization with each agent type

3.4 Drill Down Functionality

Contrary to the different preferences for 4 versus 8 COAs, participants' opinions were generally in agreement about the two methods of drilling down for information. Most preferred using the hover functionality on the Vehicle Comparison Tool that called up the specific values, listed in order from best to worst, for all 8 vehicles for the corresponding parameter (see Fig. 3). These participants commented that it was easy to jump between different parameters with this method and appreciated it giving values for all the vehicles for that parameter at the same time. The hover functionality on the individual vehicle buttons that called up all four parameters for that specific vehicle/COA was less preferable. Only a single participant preferred this method citing that it facilitated evaluating each vehicle as a whole. Experimenter observations indicated that most participants used both methods during the course of trials, probably because both required participants' recall of other pertinent information to inform their selections. Specifically, each hover function either did not provide information on other parameters (Vehicle Comparison Tool) or information on other COAs/vehicles (Vehicle Selection Buttons).

4 Summary

The present experiment confirmed the results from an earlier study examining autonomous vehicle management [5]—presenting multiple COAs was better than a single COA. Even though a more temporally demanding task environment was employed in the present experiment, both the objective and subjective data were more favorable with the 4 and 8 COA visualizations compared to data with the single COA visualization, regardless of agent comprehensiveness. It should be noted, though, that it could be that a single COA visualization would be advantageous if participants had a complex multi-task environment or if COA selection had even more severe temporal constraints. In other words, participants might have elected to trust a single agent-generated COA as best (and the only information needed) in a more challenging task environment.

The present experiment was also similar to the earlier one as there was not a definitive best option between the 4 and 8 COA visualizations. For instance, mean score did not significantly differ between the two multi-COA visualizations. While mean response time was significantly faster with 4 COAs than 8, it could be argued that 3 s is not a practical difference. There were also no noteworthy differences in the subjective data; participants' ratings were divided between the two multi-COA presentations. Some preferred the 4 COA visualization, a less cluttered presentation of the agent's four best alternatives, with additional information available via the drill down functionality. Others preferred the 8 COA visualization, stating it provided an overall comparison of a high number of alternative COAs with less need to drill down.

When given the option to choose amongst the visualizations during the supplemental trials, participants chose either 4 or 8 COAs depending on agent comprehensiveness. The participants' visualization preferences and strategies used in comparing parameters across COAs illustrate that individual differences and agent comprehensiveness need to

be considered in future interface design. One approach would be to employ a procedure similar to that examined in the supplemental trials: display the user's pre-selected default COA visualization, but also allow that view to be switched, by the user, to alternative visualizations. Indeed, the ability to rapidly switch between different visualizations (such as 4 and 8 COAs) may reduce the need to use hover functionality to call up additional information. Another approach would be to scope the visualization based upon agent comprehensiveness. Thus a system utilizing an agent with a highly comprehensive reasoning scheme would employ a more minimal visualization approach (such as 4 COAs) than a system using an agent with a limited reasoning scheme.

Lastly, this experiment demonstrated that for the human teammate to have an appropriate calibrated trust in the agent partner, more information about the agent's reasoning needs to be communicated (i.e., how the agent arrived at recommended solution(s)). Participants' comments indicated that many of them compared the result of their personal decision-making with the agents' recommendations and assessed plausible differences in their respective reasoning. This explains participants' requests for more insight into how agents utilized each COA parameter in tradeoff analyses. Control functionality that enables certain vehicles/parameters/COAs to be grouped together on the visualization might also facilitate comparisons and inform selections. Any new visualization and control functionality technique would need to be evaluated, though, to ensure that it enhances accurate and timely decision making.

Acknowledgment. This work was funded by the Air Force Research Laboratory.

References

1. Hansen, M., Calhoun, G., Douglass, S., Evans, D.: Courses of action display for multi-unmanned vehicle control: a multi-disciplinary approach. In: The 2016 AAAI Fall Symposium Series: Cross-Disciplinary Challenges for Autonomous Systems, Technical Report FS-16-03 (2016)
2. Smith, P.J.: Making brittle technologies useful (Chap. 10). In: Smith, P.J., Hoffman, R.R. (eds.) Cognitive Systems Engineering: The Future for a Changing World, pp. 181–208. CRC Press, New York (2018). https://doi.org/10.1201/9781315572529
3. Brill, E., Flach, J., Hopkins, L., Ranjithan, S.: MGA: a decision support system for complex, incompletely defined problems. IEEE Trans. Syst. Man Cybern. **20**(4), 745–757 (1990). https://doi.org/10.1109/21.105076
4. Behymer, K.J., Mersch, E.M., Ruff, H.A., Calhoun, G.L., Spriggs, S.E.: Unmanned vehicle plan comparison visualizations for effective human-autonomy teaming. Proc. Manuf. **3**, 1022–1029 (2015). https://doi.org/10.1016/j.promfg.2015.07.162
5. Behymer, K., Ruff, H., Calhoun, G., Bartik, J., Frost, E.: Presentation of autonomy-generated plans: determining ideal number and extent differ. In: Chen, J. (ed.) AHFE 2018. AISC, vol. 784, pp. 90–101. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-94346-6_9