Automated Aiding of Image Analysts as a Function of Sensor Type

Dr. Holly Zelnio and Dr. Mary Fendley
711th Human Performance Wing
Wright State University
UNITED STATES

holly.zelnio@us.af.mil

ABSTRACT

The use of an automated Decision Support System (DSS) to aid an image interpreter performing a target classification task is analyzed to ascertain performance improvement due to human automation cooperation. The performance of the image interpreter and DSS is analyzed from a human factors perspective. This research performs a human subject experiment measuring classification performance as a function of classification accuracy, interpretation time, confidence in the decision, and trust in the decision support system (DSS). In addition, multiple types of automated support were compared as were multiple types of imaging sensors. The imaging sensors that were used in this study were electro-optical (EO), infrared (IR), and synthetic aperture radar (SAR). The DSSs were developed using cognitive task analysis and previous experimentation. The results show that different types of sensors require different DSSs for optimal performance. Results suggest that EO requires less automated support while the more non-literal sensors (e.g., IR and SAR) need more automated support for highest performance. This research also investigated how a DSS itself can cause a bias influencing performance. The performance of the human and machine team can be enhanced by the presentation of key information. It was shown that the level of automated analyst aiding affects both the quality and timeliness of the decision as well as the trust in the DSS.

1.0 BACKGROUND

Trust has to be balanced for optimum performance. Either too much or too little trust can result in failures [1]. To maintain trust in automation, Sarter suggests that the human should understand the basis of the automated aids but should have the task responsibility. The human can use the automation for help, but the automation is now a means to the end of a human responsible task [2]. Increased transparency has led to increased trust, but higher decision times and workloads [3]. Increased trust and performance is suggested by findings if contextual information is provided by the automated DSS [4]. Evidence indicates that, over time, experienced workers can learn to calibrate their trust for the automated system used in their job [5].

Biases can appear when making judgments. It is difficult to fully eliminate a bias from a decision. However, two approaches to help mitigate bias includes: 1) additional training [6] or knowledge needs to be given to the individual or 2) the task needs to be structured to not create these biases. The task structuring can be designed to coincide with an individual's natural thought process to help inhibit bias [7]. This approach can be accomplished by examining where the bias is coming into play either via the person or the task/environment. Several different approaches have been studied to suppress bias including using structured processes [7], applying a three stage framework [8], improving the environment [9], and adopting the specific debiasing strategies of technical, motivational, and cognitive [10].

Cognitive bias happens when one deviates from good judgement and can occur or increase when introducing information to a decision maker. Arnott [11] has an extensive list of biases that influence people. Fendley [12] shortened this list to specifically biases that are found in object identification. The dominant biases that were identified by Zelnio and Fendley [13] when analysts perform object identification were confirmation and anchoring. Confirmation bias is present when the data and interpretation favors a single decision over alternatives [14-16]. Anchoring bias is being reliant on an initial value or given information [17, 18]. Overall, it has been suggested that a DSS should mitigate bias [19, 20]. The number of studies performed confirming bias exists heavily outweighs the number of studies that try to mitigate bias [21]. It is understood that bias cannot be eliminated. It is important to acknowledge that it exists and then try to design the support system to use the biases to our advantage or to try to make them negligible.

DSSs can be built in various ways including: a framework for the characterization of the decision situation [22], three factor function for designers to structure representation [23], the summary table of aiding requirements (STAR) table [24], and the operator function model (OFM) [25]. An OFM serves as a discrete control model for human decision making. An OFM is built through subject matter expert interviews and the Applied Cognitive Task Analysis (ACTA) process. OFM's have been used in image interpretation [26] and NASA ground control system interface [27]. The experiment in this paper builds on the image classification experimentation from Zelnio and Fendley [13].

2.0 METHODS

2.1 Participants

Thirty participants (21 male, 9 female) between the ages of 18-50, with an average age of 26.4, were recruited from the college community. Twenty-five participants answered yes to having experience with EO, IR, and SAR. Five participants responded to having knowledge of all three sensor types, but did not have experience with all three. Recruiting was conducted either through face-to-face contact or through personal e-mails. Study permission was obtained from the University IRB, and no compensation was provided.

2.2 Experiment Setup

The participant sat approximately 18 inches from the Tobii eye tracking monitor for this experiment. The different images and DSS sets were displayed on the screen, and the participants selected a response using the mouse. The participant responses and eye tracking data were collected using the Tobii program.

2.3 Stimuli

The equipment used in the study consisted of a computer display showing images drawn from publically available datasets. The Moving and Stationary Target Acquisition and Recognition (MSTAR) images consist of labeled military vehicles in SAR imagery and are available online. The electro-optical and infrared data were rendered using Meta-VR. The experiment was conducted in an access-controlled laboratory.

In the experiment, there were fifteen sets of images for the participant to classify. The images were from three sensor sources: electro-optical (EO), infrared (IR) and synthetic aperture radar (SAR). There were five sets of twelve images for each sensor. The participants classified whether there was a target in the image and if so, which target type was being shown. The targets in this experiment are military vehicles: T-72, BMP-2, and BTR-70. Four of the image sets for each sensor were displayed with an accompanying DSS. The other set did not include a decision aid to help the participant with their classification. For the four sets of images with a DSS, the participant was asked why they made that decision. They had six choices that were designed to uncover the cognitive heuristics that were being used. All sensors' images were tested in this way. All participants were asked to provide a ranking of their confidence for each decision using a 5-point Likert scale. The tests were presented in random order within each sensor.

2.4 Experimental Design

The experiment was a fixed-effect experiment and a within subject design. The experiment was divided into five alternative designs: 1) no DSS, 2) DSS consisting of a recommended classification with a corresponding percent confidence, 3) DSS consisting of a recommended classification with a corresponding percent confidence and, in addition, reference images, 4) DSS consisting of the highest and second highest recommendations and corresponding percent confidences, 5) DSS consisting of the highest and second highest recommendations and corresponding percent confidences and, also, reference images. The participants had four classification options for each test image: T-72, BMP-2, BTR-70, and No Target. After each classification, the participant was asked to state their confidence in the correctness of their decision on a five-point Likert scale. Each classification was timed. The participant was also asked why they made that decision when the DSS was present by answering the following questions:

How did you use the Decision Aid in the last decision?

1a. Confirmation of what I already thought was correct using the PERCENTAGE more
1b. Confirmation of what I already thought was correct using the IMAGES more
1c. Confirmation of what I already thought was correct using the BOTH the percentage and the images equally
2a.Used the suggestion for decision using the PERCENTAGE more
2b. Used the suggestion for decision using the IMAGES more
2c. Used the suggestion for decision using BOTH the percentage and the images equally
3. A specific part of the image influenced your decision, solely using the test image and not using the decision aid
4. None of the above

After each set of twelve images in the decision aid portion of the test was completed, a trust rating was collected.

For this study, the independent variables were the sensor types and levels of DSS. The dependent variables were time, accuracy, confidence, bias, and trust rating. The order in which the participant viewed each sensor was randomly selected. In addition, the data set order with and without the DSS was randomly selected for each participant. Within each set, the images were randomized initially, and this presentation order remained the same for each participant.

3.0 ANALYSIS

The analysis is divided into three sections corresponding to the two independent variables: the sensor level (EO, IR, SAR) and the DSS levcl (corresponding to the 5 levels described above), and the analysis of the heuristic data. For the sensor level analysis, the performance is not broken out by DSS levels; however, the DSS level analysis does break out the performance as a function of sensors and the dependent variables of accuracy, confidence, time, and trust. The use of heuristics by the image analysts considered both sensor and DSS effects. In addition, the possible biases were also investigated.

3.1 Sensor Level

Accuracy analysis: The accuracy is analyzed using the Marascuillo procedure. Accuracy is first analyzed for each sensor. It was found that each sensor's accuracy was statistically different than each other. EO had the highest accuracy ($M=88.61$). IR had the second highest accuracy ($M=77.99$), and SAR has the lowest accuracy (73.49).

Confidence analysis: Using matched pairs, the EO Confidence ($M=4.11$) was greater than IR ($t=11.3759$, $p=<.0001$) and SAR Confidence ($t=17.9958$, $p<.0001$) and the IR Confidence ($M=3.547$) was greater than the SAR Confidence ($M=2.98$) ($t=9.59552$, $p=<.0001$).

Trust Analysis: The EO trust ($M=4.04$) was greater than the IR trust ($M=3.94$) ($t=2.12043$, $p=.018$) and the SAR trust ($M=3.625$) ($t=7.04389$, $p=<.0001$). In addition, the IR trust was greater than the SAR trust ($t=5.20151$, $p=<.0001$).

Interpretation time analysis: The EO Time ($M=5.04$) to make a call was less than the IR time ($M=6.37$) and the SAR time ($M=7.828$) ($t=-5.241772$, $t=-7.4516$, both with $p=<.0001$). The IR time was less than SAR Time ($t=-4.515577$, $p=<.0001$).

Sensor level conclusion: All paired tests were significant at the $\alpha=0.05$ level. Based on the Marascuillo procedure, EO Accuracy > IR Accuracy > SAR Accuracy. The results are depicted in Figure 1 below demonstrating the relationship of EO (most literal) to IR (less literal) to SAR (least literal) to the quantitative results reported in this section. These results support the hypothesis that performance correlates with the literalness of the sensor.

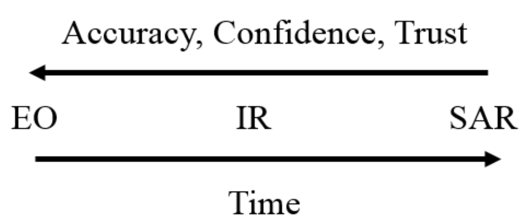


Figure 1: Spectrum of Sensor Results

3.2 DSS Level

Using the Marascuillo procedure to compare accuracy, DSS1 ($M=63.70$, no DSS) was clearly less accurate that the other DSS levels, and this difference was statistically significant. The only other statistically significant difference was between DSS 3 ($M=80.73$) and DSS 4 ($M=86.54$); however, this difference does not support any meaningful conclusion as it was grouped within the other DSS systems.

Accuracy analysis: For EO, DSS 4 ($M=96.11$) and DSS 5 ($M=95.83$) have the highest accuracy and are statistically different from the other DSS systems but were not different from each other. Both DSS 4 and 5 gave the top two confidence values. Also, for EO, DSS1 was not the poorest performing DSS system likely due to the literalness of EO. For IR and SAR, the only significant difference is between DSS 1 (IR DSS1 M=58.33 and SAR DSS1 M=45.55), the lowest accuracy system, and the other four DSS systems. The individual sensor analysis is consistent with the previous conclusion that the primary and only consistently statistically significant difference is between no DSS and any of the other DSS systems.

Confidence analysis: For all sensors, all decision aids were statistically higher in confidence than no decision aids (M=3.14). DSS 3 (M=3.74) and DSS 5 (M=3.72), the imagery examples, gave the highest confidence, and the difference between them was not statistically significant (t=.44762, p=.3278). Both were statistically greater than DSS 4 (M=3.47) (t=4.1511, p=<.0001; t=5.0696, p=<.0001 respectively). DSS 3 was statistically greater than DSS 2 (M=3.67) (t=1.7175, p=.0447), but DSS 5 was not statistically greater than DSS 2 (t=.9645, p=.1687).

For SAR confidence, DSS3 (M=3.31) was statistically significantly higher than the rest of the DSSs (for DSS1 t=8.422, p=<.0001; DSS2 t=2.0907, p=.0227; DSS4 t=3.2629, p=.0014; and DSS5 t=2.1593, p=.0199). The next highest were DSS 5 (3.11) and DSS 2 (M=3.16) which were not statistically different (t=-.70161, p=.2443). For IR, DSS 3 (M=3.92), DSS 2 (3.87), and DSS 5 (M=3.73) gave the highest confidence with DSS 3 and DSS 2 being statistically higher than DSS 5 (t=2.796, p=.0045 and t=2.0777, p=.0233). Also, DSS 3, DSS 2, and DSS 5 were statistically significantly higher than DSS 4 (M=3.11) (t=12.3566, p=<.0001; t=10.7366, p=<.0001; t=8.7364, p=<.0001). For EO, not all of the DSSs (DSS 2 (M=3.994) and DSS 3 (M=3.989)) were a statistical improvement over DSS 1 (M=3.97). However, DSS 4 (M=4.306) and DSS 5 (M=4.314) were statistically greater than the rest of the DSSs.

Trust analysis: Overall, the highest trust occurred with DSS 3 (M=4.02) and DSS 5 (M=3.88), which had imagery examples. DSS 3 was statistically significantly greater than all other DSSs whereas DSS5 was statistically greater than 4 (t=2.962, p=<.0001) but not DSS 2 (M=3.85) (t=.567653, p=.2858) at alpha=0.05.

For SAR, both DSS 3 (M=3.73) and DSS 5 (M=3.61) were again the highest trust, but the only significant difference was between DSS 3 and DSS 4 (M=3.54), DSS 4 < DSS 3 (t=-2.557, p=.008). Again, for IR, DSS 3 (M=4.14) and DSS 5 (M=3.97) gave the highest trust, but only the differences with DSS 3 were statistically significant (with DSS 2 t=3.13676, p=.0019; DSS 4 t=4.33296, p=<.0001; and DSS 5 t=2.40832, p=.0113). With EO, there was a similar story with DSS 3 (M=4.2) and DSS 5 (M=4.07), the ones with imagery examples giving the highest trust scores, again, differences of DSS 3 were statistically significant (with DSS 2 t=2.2685, p=.0155; DSS 4 t= 4.6697, p=<.0001; and DSS 5 t=1.8692, p=.0359). DSS 5 was statistically greater than DSS 4 (M=3.85) (t=2.641, p=.0066) at alpha=0.05.

Interpretation time analysis: Total time used with each DSS was ranked consistent with the amount of information presented. DSS 1 used the least amount of time (M=4.92), then DSS 2 (M=5.64) and DSS 4 (M=6.32) which were not statistically different but was the next in line of time, and DSS 3 (M=6.48) and DSS 5 (M=8.72) used the most amount of time. DSS 2 and DSS 4 showed no imagery whereas both DSS 3 and DSS 5 showed imagery and hence took the longest. Next, based on the eye tracking metrics, a deep dive was taken into where the time was spent when using the DSSs. The heat maps shown in Figure 2, 3, and 4 the duration of fixations and are a visual way of viewing fixations. It is noted that the heat maps depict the cumulative set of fixations of the 30 participants as they gazed individually at a single DSS test example. It is noted that when there are two suggestions, it appears most of the time is spent on the first or top suggestion.

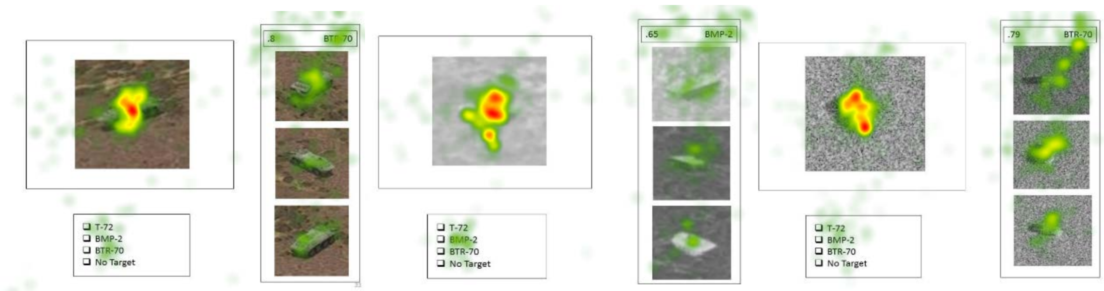


Figure 2: DSS 3 Heat Maps of Fixation Duration EO, IR, SAR

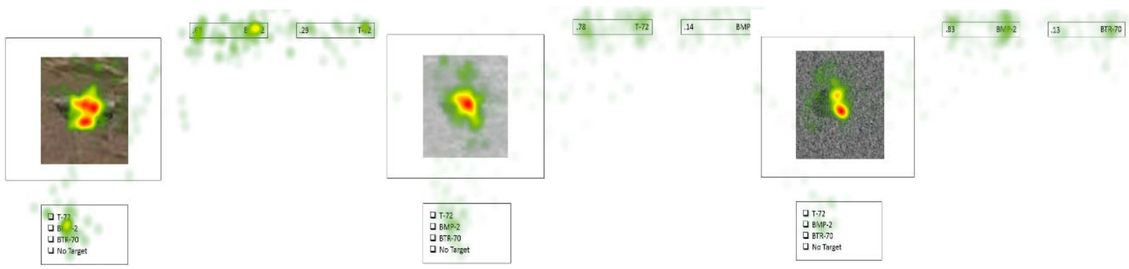


Figure 3: DSS 4 Heat Maps of Fixation Duration EO, IR, SAR

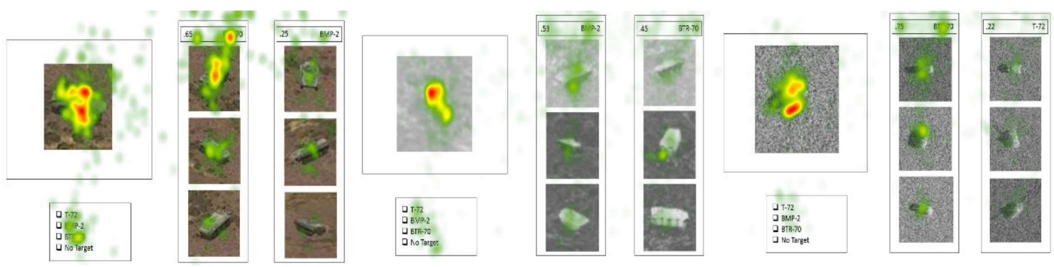


Figure 4: DSS 5 Heat Maps of Fixation Duration EO, IR, SAR

DSS level conclusion: The use of DSSs improved the confidence in the decision over no DSS provided, and these differences were statistically significant in all cases. In addition, the use of imagery in the DSS for all cases increased the trust ($t=4.655146$, $p=<.0001$) and confidence ($t=4.96363$, $p=<.0001$) over DSSs without imagery (trust with imagery $M=3.95$, trust with percentage $M=3.78$, confidence with imagery $M=3.73$, confidence with percentage $M=3.57$). It was interesting that DSS 3, the one with only one confidence value and one set of imagery examples was consistently the highest scoring DSS in both the trust and confidence category. The additional confidence and imagery information given in DSS 5 did not improve the confidence or trust overall but did slow the decision process. In fact, as the amount of information provided increased, the interpretation time increased.

3.3 Heuristic Data

We have established that cognitive heuristics are used to avoid mental workload and may have a negative impact on decision making. As discussed earlier, confirmation and anchoring were more clearly examined in

Experiment 2. For both confirmation and anchoring, there is a lack of adjustment in the decision making. For confirmation, there is no adjustment from the answer that the participant believes is true. The participant does this by only looking for confirming evidence of what they already believe is correct. When finding this confirming evidence, the affirmation can lead to higher confidence in the decision. For anchoring, there is lack of adjustment from the information given. The DSS provides a percentage and/or images that could be used by the participant. The participants could anchor on a piece or pieces of information and not adjust by using their own perception or secondary information. Overall, we are looking at the adjustment heuristic broken down into the subcategories of confirmation and anchoring. The question used to collect the heuristic and bias data with the explanation of the answers are shown in Table I.

Table 1: Heuristic/Bias Question and Explanation

Heuristic/Bias Question Used in Experiment 2	Heuristic Being Used
1a. Confirmation of what I already thought was correct using the PERCENTAGE more	Participant has made their decision and uses the percentage on the DSS to confirm their decision- Confirmation
1b. Confirmation of what I already thought was correct using the IMAGES more	Participant has made their decision and uses the images on the DSS to confirm their decision- Confirmation
1c. Confirmation of what I already thought was correct using the BOTH the percentage and the images equally	Participant has made their decision and uses the both the percentage and the images on the DSS to confirm their decision equally - Confirmation
2a. Used the suggestion for decision using the PERCENTAGE more	Participant is unsure of the answer and uses the suggestion of the DSS based off of the percentage- Anchoring
2b. Used the suggestion for decision using the IMAGES more	Participant is unsure of the answer and uses the suggestion of the DSS based off of the image- Anchoring
2c. Used the suggestion for decision using BOTH the percentage and the images equally	Participant is unsure of the answer and uses the suggestion of the DSS based off of the percentage and images equally- Anchoring
3. A specific part of the image influenced your decision, solely using the test image and not using the decision aid	Did not use DSS, no heuristic used
4. None of the above	No heuristic used

The impact of the use of cognitive heuristics was analyzed by looking at the responses by sensor type, DSS type and incorrect responses, indicating a cognitive bias.

First, the correct answer heuristic data was evaluated as a function of sensor type. The initial question addressed was whether or not the participants used various heuristics dependent upon the sensor type. Whereas, for the least literal sensor, the participants relied on the decision aid as the basis for their decision. For EO, confirmation was used 75 percent of the time when the correct answer was selected and 62 percent of the time for IR. Anchoring was used 62 percent of the time for SAR when the correct answer was selected. This indicates that, for the more literal sensors, the participants used the DSS to confirm their answer. All the differences among the three sensor types were statistically significant, as shown in Figure 5. This follows the pattern that was set for all heuristic data for all responses.

Categorical

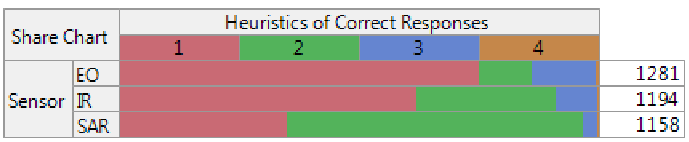
Responses(Heuristics of Correct Responses) By Sensor

Freq / Share			Heuristics of Correct Responses 1	2	3	4	Total Responses	Compare
Sensor	EO	A	958 74.8%	140 10.9%	171 13.3%	12 0.9%	1281	B,C
	IR	B	740 62.0%	343 28.7%	105 8.8%	6 0.5%	1194	C
	SAR	C	405 35.0%	712 61.5%	37 3.2%	4 0.3%	1158	

Default Comparison Groups: A/B/C

Shows letter of the category it is significantly different from at the higher share level

* Base count warning	100	Uppercase Alpha Level	0.05
** Base count minimum	30	Lowercase Alpha Level	0.1



Compare Each Sample

Letter comparisons use Pearson Chisq

Sensor, Heuristics of Correct Responses

LR PValues — LR Chi-square p-value on pairs

	A	B	C
A	1.0000	<.0001	<.0001
B	<.0001	1.0000	<.0001
C	<.0001	<.0001	1.0000

Pearson PValues — Pearson Chi-square p-value on pairs

	A	B	C
A	1.0000	<.0001	<.0001
B	<.0001	1.0000	<.0001
C	<.0001	<.0001	1.0000

Figure 5: Heuristic for Correct Responses by Sensor

The correct responses were next analyzed as a function of DSS type. The significant differences were seen between the following decision support levels: DSS 2 (confidence only) and DSS 5 (images and confidences), and DSS 4 (confidences only) and DSS 5 (images and confidences). This does follow a similar pattern with the statistical differences were between a DSS that used confidence(s) and a DSS that had both images and confidences. These differences are shown in Figure 6. Again, the differences as a function of DSS were subtle

as compared to the sensor differences. It would be hard to conclude that the DSS types had a significant effect on the heuristics that the analyst used.

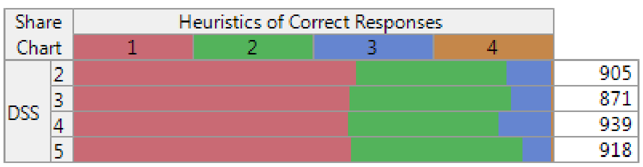
Categorical

Responses(Heuristics of Correct Responses) By DSS

Freq / Share			Heuristics of Correct Responses 1	2	3	4	Total Responses	Compare
DSS	2	A	533 58.9%	284 31.4%	82 9.1%	6 0.7%	905	
	3	B	502 57.6%	291 33.4%	73 8.4%	5 0.6%	871	
	4	C	536 57.1%	295 31.4%	103 11.0%	5 0.5%	939	D
	5	D	532 58.0%	325 35.4%	55 6.0%	6 0.7%	918	A

Default Comparison Groups: A/B/C/D

Shows letter of the category it is significantly different from at the higher share level
* Base count warning 100 Uppercase Alpha Level 0.05
** Base count minimum 30 Lowercase Alpha Level 0.1



Compare Each Sample

Letter comparisons use Pearson Chisq

DSS, Heuristics of Correct Responses

LR PValues

LR Chi-square p-value on pairs

	A	B	C	D
A	1.0000	0.8069	0.5579	0.0455
B	0.8069	1.0000	0.2928	0.2463
C	0.5579	0.2928	1.0000	0.0011
D	0.0455	0.2463	0.0011	1.0000

Pearson PValues

Pearson Chi-square p-value on pairs

	A	B	C	D
A	1.0000	0.8069	0.5588	0.0462
B	0.8069	1.0000	0.2951	0.2470
C	0.5588	0.2951	1.0000	0.0012
D	0.0462	0.2470	0.0012	1.0000

Figure 6: Heuristics for Correct Responses by DSS

Now that the use of the heuristics is established, the possible biases were analyzed by partitioning the data to contain only incorrect responses. As can be seen in Figure 7, the potential impact of the anchoring bias was demonstrated the most often when they answered incorrectly with all three sensors. The potential biases used were as follows: EO used anchoring 54 percent of the time while IR used anchoring 47 percent of the time. These differences were not found to be statistically significant; however, SAR used anchoring 63 percent of the time which was statistically different from both EO and IR.

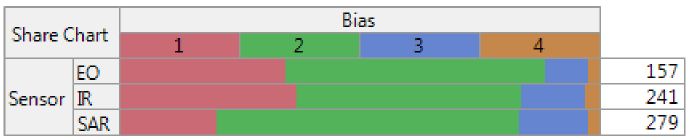
Categorical

Responses(Bias) By Sensor

Freq / Share			Bias 1	2	3	4	Total Responses	Compare
Sensor	EO	A	54 34.4%	85 54.1%	14 8.9%	4 2.5%	157	
	IR	B	89 36.9%	112 46.5%	32 13.3%	8 3.3%	241	
	SAR	C	56 20.1%	176 63.1%	40 14.3%	7 2.5%	279	A,B

Default Comparison Groups: A/B/C

Shows letter of the category it is significantly different from at the higher share level
* Base count warning 100 Uppercase Alpha Level 0.05
** Base count minimum 30 Lowercase Alpha Level 0.1



Compare Each Sample

Letter comparisons use Pearson Chisq

Sensor, Bias

LR PValues

LR Chi-square p-value on pairs

	A	B	C
A	1.0000	0.3768	0.0084
B	0.3768	1.0000	0.0002
C	0.0084	0.0002	1.0000

Pearson PValues

Pearson Chi-square p-value on pairs

	A	B	C
A	1.0000	0.3838	0.0078
B	0.3838	1.0000	0.0002
C	0.0078	0.0002	1.0000

Figure 7: Potential Bias by Sensor

When looking at the potential biases in terms of the DSSs, all DSS types had anchoring as the heuristic used most often. Secondarily, the potential biases were distributed differently among the four choices and these differences were statistically significant except for the difference between DSS 3 and DSS 5. These two DSS types both provided image examples as part of their decision systems which may explain their similar results. The statistical analysis is shown in Figure 8.

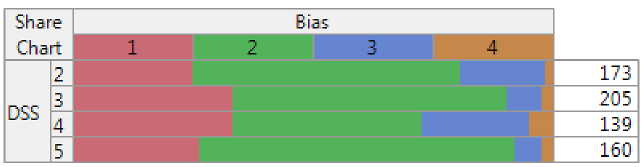
Categorical

Responses(Bias) By DSS

Freq / Share			Bias 1	2	3	4	Total Responses	Compare
DSS	2	A	43 24.9%	96 55.5%	31 17.9%	3 1.7%	173	C,D
	3	B	68 33.2%	117 57.1%	15 7.3%	5 2.4%	205	A,C
	4	C	46 33.1%	55 39.6%	31 22.3%	7 5.0%	139	
	5	D	42 26.3%	105 65.6%	9 5.6%	4 2.5%	160	C

Default Comparison Groups: A/B/C/D

Shows letter of the category it is significantly different from at the higher share level
* Base count warning 100 Uppercase Alpha Level 0.05
** Base count minimum 30 Lowercase Alpha Level 0.1



Compare Each Sample

Letter comparisons use Pearson Chisq

DSS, Bias

LR PValues

LR Chi-square p-value on pairs

	A	B	C	D
A	1.0000	0.0105	0.0255	0.0050
B	0.0105	1.0000	0.0001	0.4044
C	0.0255	0.0001	1.0000	<.0001
D	0.0050	0.4044	<.0001	1.0000

Pearson PValues

Pearson Chi-square p-value on pairs

	A	B	C	D
A	1.0000	0.0110	0.0263	0.0068
B	0.0110	1.0000	0.0001	0.4071
C	0.0263	0.0001	1.0000	<.0001
D	0.0068	0.4071	<.0001	1.0000

Figure 8: Potential Bias by DSS

4.0 DISCUSSION

For classification accuracy, all the DSSs that were designed and tested outperformed the no DSS conditions. However, the interpretation time increased with additional information in the DSS. This finding suggests that for time critical tasks, careful attention must be paid to the amount of information in the decision aid as a more accurate decision that is not made within the necessary timeline would not meet mission objectives. From a sensor perspective, the classification accuracy of EO was greater than IR which, in turn, was greater than SAR which was in order of the literalness of the sensor. As may be anticipated, the timing numbers were reversed with SAR taking longer than IR which took longer than EO.

Both trust and confidence were enhanced with the DSSs; however, these subjective measures and eye tracking data showed that the reference imagery was the major contributor to the participant's trust and confidence, not the DSS algorithm probability. The study dug deeper in attempting to ascertain the dependency of the interpretation task on sensor type. To accomplish this goal, the heuristics that the participants used to make their decisions were elicited. It was found that for EO and IR, confirmation (decision primarily driven by test imagery) was the heuristic that was used most often, and for SAR, anchoring (decision primarily driven by DSS) was used most often. Again, this is understandable based on the literalness of each sensor. Also, this finding was also supported by the eye tracking data as the fixation time on the reference imagery increased with the non-literalness of the sensor (i.e., SAR time > IR time > EO time). On the other hand, when considering possible biases (i.e., errors resulting from heuristic use), anchoring was the culprit for all sensor types.

5.0 FUTURE WORK

This experiment can be further analyzed by looking into the fixation count, visit count, and transitions of areas of interest within the DSS. The eye tracking data can be used to see if there are any other lessons learned, specifically, exploring the effect of the secondary suggestion. Did the secondary suggestion have a positive impact or was it too much information? Eye tracking will also be able to give a firm answer to whether the secondary suggestion was considered by determining if the participants looked at this area of the DSS.

REFERENCES

[1] Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5), 527-539.

[2] Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of human factors and ergonomics*, 2, 1926-1943.

[3] Helldin, T., Falkman, G., Riveiro, M., Dahlbom, A., & Lebram, M. (2013). Transparency of military threat evaluation through visualizing uncertainty and system rationale. *Engineering Psychology and Cognitive Ergonomics*. Applications and Services, 8020, 263-272.

[4] Liebhaber, M. J., & Feher, B. (2002). Air threat assessment: Research, model, and display guidelines. SPACE AND NAVAL WARFARE SYSTEMS COMMAND SAN DIEGO CA.

[5] Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *Intelligent Systems*, IEEE, 28(1), 84-88.

[6] Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2372732215600886.

[7] Fischoff, B. (1981). *Debiasing* (No. PTR-1092-81-3). DECISION RESEARCH EUGENE OR.

[8] Keren, G.; Caverni, J.P.; Fabre, J.M.; et al. *Cognitive Biases.* Amsterdam: North-Holland 1990.

[9] Soll, J. B., Milkman, K. L., & Payne, J. W. (2014). A user's guide to debiasing.

[10] Larrick R. P. (2004). "Debiasing," in *Blackwell Handbook of Judgement and Decision Making* eds Koehler D. J., Harvey N., editors. (Malden, MA: Blackwell Publishing Ltd.) 316–337.

[11] Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach.

Information Systems Journal, *16*(1), 55-78.

[12] Fendley, M., & Narayanan, S. (2012). Decision aiding to overcome biases in object identification. *Advances in Human-Computer Interaction*, *2012*, 7.

[13] Zelnio, H., & Fendley, M. (2018). Human performance modelling for image analyst decision support design. *International Journal of Human Factors Modelling and Simulation*, *6*(2-3), 184-202.

[14] Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes*, *66*(1), 102-110.

[15] Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.

[16] Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

[17] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

[18] Wall, E., Blaha, L. M., Paul, C. L., Cook, K., & Endert, A. (2017). Four perspectives on human bias in visual analytics. In *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations*.

[19] George, J. F., Duffy, K., & Ahuja, M. (2000). Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems*,*29*(2), 195-206

[20] Bhandari, G., Hassanein, K., & Deaves, R. (2008). Debiasing investors with decision support systems: An experimental investigation. *Decision Support Systems*, *46*(1), 399-410.

[21] Zenker, F., Dahlman, C., & Sarwar, F. (2015). Reliable Debiasing Techniques in Legal Contexts? Weak Signals from a Darker Corner of the Social Science Universe. *The psychology of argument: cognitive approaches to argumentation and persuasion*.

[22] Zachary, W. W. (1980). *Decision aids for naval air ASW* (No. 1366-A). ANALYTICS INC WILLOW GROVE PA.

[23] Woods, D. D., & Roth, E. M. (1988). Cognitive Systems Engineering. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 3-43): Elsevier Science.

[24] Hopson, J., Zachary, W., & Lane, N. (1981). The intelligent use of intelligent systems: Problems in engineering man/machine symbiosis. *AGARD The Impact of New Guidance and Control Systems on Mil. Aircraft Cockpit Design 16 p(SEE N 82-13048 04-01)*.

[25] Mitchell, C.M. & Miller, R.A. (1986) A discrete control model of operator function: A methodology for information display design, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 16 No.3, pp.343-357.

[26] Muller, Amanda C., & S. Narayanan. Cognitively-engineered multisensor image fusion for military

applications. *Information Fusion* 10.2 (2009): 137-149.

[27] Mitchell, Christine M. "GT-MSOCC: A domain for research on human-computer interaction and decision aiding in supervisory control systems." Systems, Man and Cybernetics, IEEE Transactions on 17.4 (1987): 553-572.