Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 4, 2023
A Regional Conference of the Society for Industrial and Systems Engineering

An Analytic Framework for Assessing Artificial Intelligence and Assistive Automation Enabled Command and Control Decision Aids for Mission Effectiveness

Thomas Mitchell[1], Noah Sheffield[1], Darius Richardson[1], Benjamin Jensen[1], Emily Nack[2], Iain Cruickshank[2], Robert Thomson[2], and Nathaniel D. Bastian[1,2]

[1]Department of Systems Engineering,
United States Military Academy,
West Point, New York 10996

[2]Army Cyber Institute,
United States Military Academy,
West Point, New York 10996

Corresponding author's Email: nathaniel.bastian@westpoint.edu

Author Note: This work is supported by the United States Army Combat Capabilities Development Command (DEVCOM) Analysis Center (DAC) under Support Agreement No. USMA 2255. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, U.S. Army, U.S. Department of Defense, or U.S. Government. The authors would like to thank the DAC Combat Simulations team for their support throughout the project.

Abstract: The U.S. Army has significant interest in operationalizing Artificial Intelligence and Assistive Automation (AI/AA) technologies on the battlefield to help collate, classify, and clarify multiple streams of situational and sensor data to provide a Commander with a clear, accurate operating picture to enable rapid and appropriate decision-making. This paper offers a methodology integrated with combat simulation output data into an analytic assessment framework. This framework helps assess AI/AA enabled Decision Aids for command and control with respect to mission effectiveness. Our methodology is demonstrated via a real-world operational vignette of an AI/AA-augmented Battalion assigned to clearing a sector of the battlefield. Results indicate that the simulated scenario with an AI/AA advantage modeled led to a higher expected mission effectiveness score.

Keywords: Assessment Framework, Analytic Hierarchy Process, Combat Simulation, Artificial Intelligence, Decision Aids

1. Introduction

The U.S. Army is currently developing Decision Aids that incorporate Artificial Intelligence and Assistive Automation (AI/AA) technologies into the operational battle space. According to the U.S. Army Maneuver Center, soldiers can be up to 10 times more effective in combat when assisted by AI/AA systems such as Decision Aids (Aliotta, 2022). A Decision Aid is a tool designed to assist Commanders in combat scenarios by reducing their decision time while improving decision quality and mission effectiveness (Shaneman, George, & Busart, 2022); these tools help collate operational data streams to assist Commanders with battlefield sense-making to help them make informed, real-time decisions. One problem associated with using AI/AA enabled Decision Aids is that the Army currently lacks a validated framework to assess tool usage in an operational environment. As such, in this paper we describe our research, design, and development of an analytic framework coupled with modeling and simulation to assess AI/AA Decision Aids for command and control in terms of mission effectiveness.

As part of our analytic framework development, we conducted extensive literature review along with stakeholder analysis with over 30 stakeholders who are knowledgeable in the domains of AI/AA, Decision Aids, command and control, and modeling and simulation. These stakeholders were placed into focus groups based on their familiarity with aforementioned topics. We conducted virtual focus group meetings with each group, gathered feedback, and used it to drive our findings, conclusions, and recommendations (FCR). Concurrently, we developed a realistic battlefield vignette and scenario. Using this scenario and our FCR output, we collaborated with the U.S. Army DEVCOM Analysis Center (DAC) to develop a functional hierarchy of objec- tives to measure through modeling and simulation. We transferred our hypothetical combat scenario into One Semi-Automated Forces (OneSAF), a simulation software that utilizes computer-generated forces, offering models of entities and behaviors that are partially or entirely automated, and intended to support Army readiness (PEOSTRI, 2023). Using the

Analytical Hierarchy Process, we elicited assessment decision-maker preferences and computed weights to objectives in the functional hierarchy and created a spreadsheet model that incorporates output data from OneSAF and provides a quantitative value score. Using A-B testing, we gathered scores for a baseline simulation as well as one in which AI/AA effects were modeled. We compared results of the A and B scenarios and assessed the effects that AI/AA had on mission effectiveness of friendly forces in the simulation.

2. Literature Review

Analytic assessment frameworks enable quantitative and/or qualitative data to be evaluated for a multiple criteria decision problem. The qualitative frameworks such as the Kano Model (Violante & Vezzetti, 2017), French Question and Answering (Hordyk & Carruthers, 2018) and Qualitative Spatial Management (Pascoe, Bustamante, Wilcox, & Gibbs, 2009) are used mainly for stakeholder input and brainstorming (Srivastava & Thomson, 2009) without intensive calculation or labor. Quantita- tive assessment frameworks are data-driven and provide a mathematical methodology to determine a system's functions through measures of performance and measures of effectiveness. The Analytic Hierarchy Process (AHP) is applicable to our problem given its use of hierarchical design with pairwise decision maker preference comparison to provide qualitative and quantitative analysis through comparative weighting (Saaty, 1987). While AHP has been used in many applications, to our knowledge this methodology has not been used to assess AI/AA enabled Decision Aids or coupled with A-B testing for assessment.

Command and control (C2) systems are used to provide a more detailed, accurate, common operating picture of the battlefield in order to enable effective decision-making; these C2 systems are largely built to increase situational awareness (SA). Studies have shown that Commanders using digitized information display methods, something an AI/AA enabled Decision Aid could enhance, display greater levels of SA than Commanders using radio communications to gather information (McGuinness & Ebbage, 2002). The value gained from AI/AA integration with C2 can be likened to a "cheat" in a combat video game: it provides an information advantage on how the enemy operates and helps friendly forces avoid costly consequences (McKeon, 2022). Research on C2 systems and SA have helped drive the development of the vignette and scenario described herein.

Modeling and simulation (M&S) is a simplified representation of a system or process that allows us to make predictions or understand the behavior through simulations. M&S generates data that allows one to make decisions and predictions based off certain scenarios (TechTarget, 2017). This allows the Army to generate and draw conclusions from operational scenarios that have been experienced and ones that the Army expects to face in the future. Simulations help drive the Army's capability assessment. Testing and evaluation often takes place alongside assessment and consists of analyzing models to learn, improve, and draw conclusions from, while also assessing risk. There are many different M&S tools used throughout the military. For example, the Infantry Warrior Simulation (IWARS) is a combat simulation focused on individual and small unit forces to assess operational effectiveness (USMA, 2023). The Advanced Framework for Simulation, Integration and Modeling (AFSIM) is a multi-domain M&S framework for simulation focused on analysis, experimentation, and wargaming (West & Birkmire, 2020). Within the scope of our project, One Semi-Automated Force (OneSAF) is used to model combat situations we have created in order to simulate the effects of having AI/AA advantages on the battlefield.

As mentioned, the goal of AI/AA-enabled Decision Aids is to increase quality and speed of decision-making. AI can be utilized for different scenarios and it can provide support to battlefield Commanders and warriors in multiple ways. For example, AI/AA enabled Decision Aids can help warriors in both air and ground combat be able to "analyze the environment" better and "detect and analyze targets" (Adams, 2001). AI/AA enabled Decision Aids can help mitigate human error and create information and decision advantage on the battlefield (Cobb, Jalaian, Bastian, & Russell, 2021). These example information triage advantages gained by AI/AA enabled Decision Aids guided our operational vignette and M&S scenario development.

3. Methodology

3.1. Operational Vignette and Scenario Development

In our operational vignette, 1st Battalion is assigned with a small village up to a designated line of advance. The vignette follows Captain Roy, the Battalion Intelligence Officer (BN S2), as he prepares the intelligence situational template (SITTEMP) using an AI/AA enabled Decision Aid (i.e., assistant) which rapidly collects and incorporates accumulated Red intelligence and open source intelligence-derived situational data. It then follows Major Jones and Captain Smith, the Battalion Operations Offi- cer (BN S3) and the Assistant S3 (AS3), as they develop maneuver courses of actions (COA) using the AI/AA enabled Decision Aids to evaluate "what-if" scenarios Finally, it switches to Lieutenant Kim, the Battalion Assistant S2 (BN AS2), as she devel- ops named areas of interest (NAI) based on the selected maneuver scheme and then works to coordinate adequate Intelligence, Surveillance, and Reconnaissance (ISR) coverage between her internal assets and upper echelon resources. Assumptions made as part of the vignette include that the time period is 2030, neither side will use nuclear weapons or take action that represents an existential threat to the other, weather conditions affect BLUE and RED forces equally, the time of the year is fall season with warm and humid weather.

3.2. Stakeholder Analysis and Functional Hierarchy Development

As part of background research for solution framing, we engaged with 32 civilian and military stakeholders who are experts in AI/AA and its contributions to decision-making and simulation-based modeling. The stakeholder analysis process we conducted is as follows: 1) Define and Identify Stakeholders; 2) Define Focus Groups; 3) Assign Stakeholders to Focus Groups; 4) Develop Questions Specific to each Focus Group; 5) Contact Stakeholders and Schedule Focus Group Sessions; 6) Conduct Focus Group Sessions; 7) Synthesize and Analyze Stakeholder Feedback; and 8) Develop FCR matrices. We used the results of the FCR matrices to develop a functional hierarchy diagram of the objectives, measures and metrics to generate/collect from the simulated scenarios. These objectives, measures and metrics were then ranked against each other in terms of importance to the mission set. This set the foundation for using the Analytic Hierarchy Process (described below).

3.3. Analytic Hierarchy Process and A-B Testing

The AHP is a methodology, originally proposed by Thomas Saaty in 1987, that utilizes a series of pairwise comparisons derived from experts' judgment that places each function and sub-function from a functional hierarchy into a prioritized scale. The various attributes are then ranked against each other through tangible data or qualitative opinions of experts. These rankings are then placed on a scale of 1-9 as seen in Table 1. After each attribute is given its weight 1-9, the criteria and sub-criteria are given weights that demonstrate their relative importance (Saaty, 1987).

Table 1. AHP Relative Ranking Scale

Scale Value	Explanation
1	Equally preferred (or important)
3	Slightly more preferred (or important
5	Strongly more preferred (or important)
7	Very strongly more preferred (or important
9	Extremely more preferred (or important)
2,4,6,8	Used to reflect compromise between scale values

Once these initial pairwise comparisons are complete, there is a series of four axioms that govern the AHP. These axioms state that given two sub-criteria, A_i, and A_j, the expert can give a preference judgment denoted by θ_{ij}. The preferences share an inverse relationship such that $\theta_{ij} = 1/\theta_{ji}$. Further, when comparing two criteria, A_i, can never be infinitely more preferred than A_j, such that $\theta_{ij} \neq \infty$. Finally, all impactful decisions in the problem can, and should, be formulated using a hierarchy.

After the sub-criteria (or alternatives) are ranked through pairwise comparison, the eigenvector method is used to compute the relative values and weights of these criteria. Equation 1 for this method is proposed by Saaty, and is computed as follows:

$$\theta v = \lambda_{max} \tag{1}$$

In Equation 2, v is the vector of relative values and λ_{max} is the maximum eigenvalue. Furthermore, by raising the matrix θ to the power of k and normalizing the result, the principal eigenvector can be determined.

$$v = \lim_{n \to \infty} \frac{\theta^k e}{e^T \theta^k e} \tag{2}$$

In the case, $e^T = (1, 1\ldots,1,1)$. The v vector is then normalized to the w vector, where $\sum_{i=1}^{n} w_i = 1$. Once the w vector is determined, λ_{max} is determined in Equation 3:

$$\lambda_{max} = \frac{\sum_{j=1}^{n} \theta_{1j} w_j}{w_1} \tag{3}$$

This totals up the ranking scores of each relative metric, and then this sum is divided by the total sum of all the metric scores added together. This achieves a relative weight for each criterion and sub-criteria. Once the AHP determines the weights, the sub-criteria weights are multiplied by the relative weights (or eigenvalues) for each criteria to get a localized weight. This calculates a globalized score that represents what each sub-criteria contributes to the scenario. The sub-criteria scores should add up to the relative weights for the main criteria they fall under. When multiple decision makers are involved, one may take geometric mean of the individual evaluations at each level (Saaty, 1987)

Our methodology also includes A-B testing to compare Scenario A (without AI/AA) with Scenario B (with AI/AA) to assess the impact of the Decision Aid on C2 mission effectiveness. A-B testing was originally designed for web traffic control where two variants of a product undergo statistical analysis in order to determine the best version (King, Churchill, & Tan, 2017).

3.4. Modeling and Simulation

OneSAF is a tool for modeling and simulating real/future operational scenarios. Our goal in utilizing OneSAF is to make the model as similar to our vignette as possible. We created SITTEMPs for BLUE/RED and used them to input desired entities into the OneSAF simulation. Once the entities were placed, we created actions and maneuvers for each entity to reflect Major Jones and Captain Roy's roles. After emplacing the RED entities we identified the area of interest, the village, Lieutenant Kim's role. Then, we setup two distinct scenarios: A (no AI/AA advantage) and B (AI/AA advantage). In Scenario A, RED is occupying a village and BLUE is set to clear it. RED is set to defend their current battle position and employ defensive measures through direct and indirect fires. The first phase of actions input for BLUE is for Alpha and Charlie Company to move tactically to their battle positions, establishing security and preparing to support Bravo Company. In the next phase, Bravo moves tactically while the Headquarters Company follows. Once Bravo is in position, they begin seizing the objective while Alpha and Charlie clear it. Alpha and Charlie's main goal is to assist Bravo and minimize BLUE casualties. Once Bravo finishes seizing and clearing the objective, the scenario is over.

To emulate the AI/AA enabled Decision Aid capability within Scenario B, we expanded on Scenario A by introducing new actions and adjusting existing settings. We provided BLUE with up-to-date information on enemy movements, terrain, and other factors that impact their ability to move, simulating the ability to make more informed decisions and move more quickly. Specifically, we increased the movement speed of BLUE from 4.15 km/hour to 88.99 km/hour. BLUE actors could now move at a speed anywhere between 0 and 88.99 km/hour. In addition to this, we introduced a new action for Alpha company to perform reconnaissance, which provides BLUE with real-time data, situational awareness, target identification, and threat detection, much like an AI/AA enabled Decision Aid would.

With increased insights and the transfer of real-time data into BLUE's decision-making process, an additional action was added, for increased support from Charlie as Alpha performed recon. By identifying potential targets and threats during recon, Charlie can take proactive measures to avoid or neutralize these threats. After incorporating these modifications into Scenario B, we were able to create a scenario that reflects an AI/AA advantage and provides BLUE with the tools and insights needed to make more informed decisions and succeed on the battlefield. To accurately assess the impacts of AI/AA capabilities within Scenario B, in comparison to Scenario A, we used OneSAF's Web Replication Tool (WRT) and Data Collection Specification Tool to run the scenarios multiple times and collect data for analysis.

Using the WRT, we ran both Scenario A and B simulations 30 times and exported the data to CSV file format. After analyzing the output data, we determined what best lined up with our predetermined metrics as detailed earlier. Unfortunately, we were not able to measure all metrics, due to a lack of data produced within the simulations. Based off the data we were able to collect, we measured the number of BLUE losses, the time before detection by RED, the time to reach mission goals, the number of RED kills, the time to locate RED, and the number of shots versus hits. For metrics we were unable to collect data from, we made the values congruent across both scenarios with respect to the mean, 95th and 5th percentile data bins.

4. Results and Discussion

Using the AHP and stakeholder preferences for our objectives, we created a spreadsheet model to analyze the simulation output data for both scenarios. The model takes stakeholder input to determine global weights for criteria and sub-criteria. The model then pulls raw data from the OneSAF simulation and converts it to usable data from normalizing scales that give the data a score 1-100. The data used in this analysis is the mean, median, 95th percentile, and 5th percentile values. This data is multiplied by its respective weight to produce a global score for each objective and sub-objective. These values are then summed to make a final mission effectiveness score to quantify how well the forces performed in the simulation. The model is applied to both scenarios within A-B testing, one with AI/AA effects incorporated into the simulation and one without. The data that was collected is seen in Table 2. For metrics where data was not collected, we made the values congruent across scenario A and B with respect to the statistical data bins. Therefore, there was no variability across scenario A and B, but there was variability between each statistical bin to simulate inherent variability in the data. The eight scores from the two simulations are depicted in Table 3, which are compared to determine the impact of the AI/AA effects on friendly force mission effectiveness.

Table 2. Results of the Statistical Analysis on the 30 Simulation Iterations

	Scenario A				**Scenario B**			
AHP Metric / Simulation Statistic	Mean	Median	95th	5th	Mean	Median	95th	5th
Number of Friendly Losses	238	237	180	303	233	238	172	301
Time Before Detection (min.)	580	578	637	528	569	567	615	530
Time to Missions Goals (min.)	359	365	328	377	353	347	325	378
Number of Enemy Killed	112	108	136	94	106	106	136	85
Time to Detect Enemy (min.)	451	453	421	468	449	451	418	471
Sensor to Shooter Time (min.)	11	2.99	0.54	38	7.7	2.86	0.50	21.8

For each of the six AHP metrics displayed in Table 2, several statistics (mean, median, 95th and 5th percentiles) from the 30 simulation iterations are provided for both scenarios, allowing us to see the variability of the simulation output data. Overall, these results of the A/B testing indicate that the AI/AA enabled Decision Aid (modeled in Scenario B) generally had a positive impact on C2 mission effectiveness. For example, the mean, 95th percentile and 5th percentile values for the number of friendly losses was lower in Scenario B compared to Scenario A. As another example, the sensor to shooter time was lower in Scenario B for all simulation statistic values. Notably, the AI/AA enabled Decision Aid did not increase the number of enemy killed.

Next, we plugged these simulation statistic values for each scenario into the AHP scales to convert into values from 1-100, allowing us to then compute the mission effectiveness score for each scenario. These scores are represented in Table 3.

Table 3. Results of the AHP

Statistic	Scenario A	Scenario B
Mean	39.627	43.748
Median	39.850	39.668
95th	82.625	84.570
5th	9.790	11.578

While the mean, 95th and 5th percentile mission effectiveness scores are higher when AI/AA is modeled in Scenario B, there are nearly negligible results in the median. It is clear to see that Scenario A performs worse on average and has more variability in the bounds. Scenario B, on the other hand, has a more compact spread with greater results on average. There is not a huge difference in these results, but our experimentation does indicate that AI/AA enabled Decision Aids generally improved C2 mission effectiveness for our operational vignette.

A potential reason for these results is based in the nature of the simulations and grading scale. Metrics, such as time before detection, are values that should be maximized. However, since Scenario B speeds up the BLUE force, this inherently reduces the time before detection, giving Scenario B a lower score (although the speed increase makes BLUE more lethal). The same principal is true for kills, which should be maximized. Since the BLUE force is faster and spends less time on the objective, there is less time to shoot, so naturally the amount of kills will drop. This dynamic is somewhat counter-intuitive in nature, but it helps explain why Scenario B mission effectiveness scores are not strictly better across all simulation statistics.

5. Conclusions: Summary, Limitations, and Future Work

In this work we demonstrated a novel methodology that serves as an analytic framework to assess AI/AA enabled Decision Aids for command and control in terms of mission effectiveness. By developing an operational vignette and subsequent scenario through modeling and simulation, and then leveraging the simulation output data in the Analytic Hierarchy Process for A- B testing, we demonstrated how AI/AA enabled Decision Aids can enhance friendly force capabilities in combat. The main limitation of this research stems from limited capabilities within OneSAF for modeling and simulation that accurately represents the effects of having an AI/AA enabled Decision Aid modeled within the scenario. For example, OneSAF does not easily support the integration of external-to-software algorithms via a software development kit. This makes it very challenging to integrate an actual AI/AA algorithm into the modeling and simulation environment to enable within-simulation inference needed to modeling emergent behaviors/actions. Moreover, OneSAF did not have the proper actions/systems in place to produce outputs for some of the measures/metrics that we needed for complete assessment. Note that we examined the simulation output data from OneSAF for each measure/metric and then categorized each measure/metric output as reliable, somewhat reliable, or unreliable. We chose to only use data for measures/metrics we classified as reliable for the AHP. Thus, our tool was not fully utilized and the mission effectiveness scores of the A-B tests were affected by a lack of data. Future work will expand on this methodology by exploring OneSAF deeper to find more simulation actions/systems, adjusting the current measures/metrics to other vignettes/scenarios, applying other multiple criteria decision analysis techniques other than AHP for comparison, developing a more enhanced analytic tool (rather than using a spreadsheet), and investigating ways to better model AI/AA effects within the simulation.

6. References

Adams, T. K. (2011). Future warfare and the decline of human decisionmaking. *The US Army War College Quarterly: Parameters*, *41*(4), 1.

Aliotta, J. (2022). *Army, west point hit milestone with robotics project.* U.S. Army. Retrieved from https://www.army.mil/article/254202/army_west_point_hit_milestone_with_robotics_project

Center, D. T. I. (2013). Afsim: The air force research laboratory's approach to making M&S ubiquitous in the weapon system concept development process. *CSIAC Journal*, *1*(4). doi: 10.21474/CSIAJ.2013.01.04.05

Cobb, A. D., Jalaian, B. A., Bastian, N. D., & Russell, S. (2021, December). Robust decision-making in the internet of battlefield things using bayesian neural networks. In *2021 Winter Simulation Conference (WSC)* (pp. 1-12). IEEE.

Hordyk, A. R., & Carruthers, T. R. (2018). A quantitative evaluation of a qualitative risk assessment framework: Examining the assumptions and predictions of the Productivity Susceptibility Analysis (PSA). *PloS one*, *13*(6), e0198298.

King, R., Churchill, E. F., & Tan, C. (2017). *Designing with data: Improving the user experience with a/b testing*. O'Reilly Media, Inc.

Letham, B., & Bakshy, E. (2019). Bayesian optimization for policy search via online offline experimentation. *J. Mach. Learn. Res.*, *20*, 145–1.

McGuinness, B., & Ebbage, L. (2002). *Assessing human factors in command and control: Workload and situational awareness metrics*. Defense Technical Information Center.

McKeon, A. (2022). *Can artificial intelligence apply gaming to military strategy? Northrop Grumman.* Retrieved from https://www.northropgrumman.com/what-we-do/can-artificial-intelligence-apply-gaming-to-military-strategy/

Pascoe, S., Bustamante, R., Wilcox, C., & Gibbs, M. (2009). Spatial fisheries management: a framework for multi-objective qualitative assessment. *Ocean & Coastal Management*, *52*(2), 130-138.

PEOSTRI. (2023). *One semi-automated forces.* Retrieved from https://www.peostri.army.mil/onesaf

Saaty, R. (1987). The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling*, *9*(3–5), 161–176. doi: 10.1016/0270-0255(87)90473-8

Shaneman, S., George, J., & Busart, C. (2022, June). Scaling distributed artificial intelligence/machine learning for decision dominance in all-domain operations. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV* (Vol. 12113, pp. 19-32). SPIE. Srivastava, A., & Thomson, S. (2009). *Framework analysis: a qualitative methodology for applied policy research.*

TechTarget. (2017). *Modeling and simulation (m&s).* Retrieved from https://www.techtarget.com/whatis/ definition/modeling-and-simulation-MS

USMA. (2023). *Combat simulation lab.* Retrieved from https://www.westpoint.edu/academics/academic-departments/systems -engineering/combat-simulation-lab

Violante, M. G., & Vezzetti, E. (2017). Kano qualitative vs quantitative approaches: An assessment framework for products attributes analysis. *Computers in industry*, *86*, 15–25.

West, T., & Birkmire, B. (2020). *Afsim: The air force research laboratory's approach to making m&s ubiquitous in the weapon system concept development process – csiac.* Retrieved from https://csiac.org/articles/ afsim-the-air-force-research-laboratorys-approach-to-making-ms-ubiquitous-in-the-weapon-system-concept-development-process/